EN.553.420
Introduction to Probability

An Exercise in Masochism

SAM DAWLEY

Johns Hopkins University

Contents

1 Foreword **1**

2 Basics of Probability **2**
2.1 Kolmogorov Axioms of Probability . . . 2
2.2 Sample Spaces . . . 2
2.3 DeMorgan's Laws . . . 2
2.4 Boole Inequalities . . . 2
2.5 Conditional Probabilities . . . 2
2.6 Law of Total Probability . . . 2
2.7 Bayes' Theorem . . . 2
2.8 Independence of Events . . . 2

3 Basics of Counting **3**
3.1 Basic Counting Principle . . . 3
3.2 Bijective Counting . . . 3
3.3 Partitioning Sets . . . 3
3.4 Partitions of an Integer . . . 3

4 Random Variables **4**
4.1 Discrete Random Variables . . . 4
4.1.1 Probability Mass Function . . . 4
4.1.2 Cumulative Distribution . . . 5
4.2 Continuous Random Variables . . . 5
4.2.1 Probability Density Function . . . 5
4.2.2 Cumulative Distribution . . . 7
4.3 Independent and Identically Distributed Random Variables . . . 7
4.4 Expected Value . . . 7
4.4.1 Law of the Unconscious Statistician (LOTUS) . . . 9
4.4.2 Linearity of the Expected Value . . . 10
4.4.3 When the Expected Value Does Not Exist . . . 13
4.4.4 Expectation When the Random Variable is Nonnegative . . . 14
4.5 Variance . . . 14
4.5.1 "LOTUS" for the Variance . . . 15
4.5.2 Nonnegativity of the Variance . . . 15
4.5.3 Final Remarks . . . 15
4.6 Central Limit Theorem . . . 16
4.6.1 The Continuity Correction . . . 17
4.7 Moment Generating Functions . . . 18
4.7.1 Continuity Theorem of Moment Generating Functions . . . 21
4.8 Indicator Functions . . . 21

5 Discrete Random Variables **23**
5.1 Discrete Uniform Distribution . . . 23
5.2 Bernoulli Distribution . . . 23
5.2.1 Probability Mass Function . . . 23
5.2.2 Cumulative Distribution Function . . . 23
5.2.3 Expected Value . . . 23
5.2.4 Variance . . . 24
5.3 Binomial Distribution . . . 25

5.3.1 Probability Mass Function . . . 25
5.3.2 Expected Value . . . 26
5.3.3 Variance . . . 28
5.4 Negative Binomial Distribution . . . 29
5.4.1 Probability Mass Function . . . 29
5.4.2 Expected Value . . . 30
5.4.3 Variance . . . 30
5.5 Multinomial Distribution . . . 31
5.5.1 Probability Mass Function . . . 31
5.5.2 Expected Value . . . 32
5.5.3 Variance . . . 32
5.6 Geometric Distribution . . . 33
5.6.1 Probability Mass Function . . . 33
5.6.2 Cumulative Distribution Function . . . 34
5.6.3 Expected Value . . . 34
5.6.4 Variance . . . 35
5.7 Hypergeometric Distribution . . . 37
5.7.1 Probability Mass Function . . . 37
5.7.2 Binomial Approximation to the Hypergeometric Distribution . . . 37
5.7.3 Expected Value . . . 38
5.7.4 Variance . . . 39
5.8 Poisson Distribution . . . 40
5.8.1 Probability Mass Function . . . 40
5.8.2 Poisson Limit Theorem . . . 41
5.8.3 Expected Value . . . 43
5.8.4 Variance . . . 43
5.9 Logarithmic Distribution . . . 45
5.9.1 Probability Mass Function . . . 45
5.9.2 Expected Value . . . 45
5.9.3 Variance . . . 45

6 Continuous Random Variables **46**
6.1 Continuous Uniform Distribution . . . 46
6.1.1 Probability Density Function . . . 46
6.1.2 Cumulative Distribution Function . . . 47
6.1.3 Expected Value . . . 47
6.1.4 Variance . . . 48
6.1.5 Moment Generating Function . . . 48
6.2 Exponential Distribution . . . 49
6.2.1 Probability Density Function . . . 49
6.2.2 Cumulative Distribution Function . . . 49
6.2.3 Expected Value . . . 49
6.2.4 Variance . . . 50
6.2.5 Moment Generating Function . . . 50
6.2.6 Memorylessness Property . . . 50
6.3 Normal (Gaussian) Distribution . . . 51
6.3.1 Probability Density Function . . . 51
6.3.2 Cumulative Distribution Function . . . 53
6.3.3 Expected Value . . . 53
6.3.4 Variance . . . 53

6.4 Log-Normal Distribution . . . 54
6.4.1 Probability Density Function . . . 54
6.4.2 Expected Value . . . 54
6.4.3 Variance . . . 54
6.5 Gamma Distribution . . . 55
6.5.1 Gamma Function . . . 55
6.5.2 Probability Density Function . . . 56
6.5.3 Expected Value . . . 57
6.5.4 Variance . . . 58
6.5.5 Moment Generating Function . . . 58
6.5.6 Gamma Distribution Connection to Other Distributions . . . 59
6.6 Beta Distribution . . . 60
6.6.1 Probability Density Function . . . 60
6.6.2 Expected Value . . . 60
6.6.3 Variance . . . 61
6.7 Other Distributions I Think Are Cool . . . 62
6.7.1 Cauchy Distribution . . . 62
6.7.2 Weibull Distribution . . . 62
6.7.3 Rayleigh Distribution . . . 62
6.7.4 Pareto Distribution . . . 62
6.7.5 Chi-Squared Distribution . . . 62
6.7.6 F-Distribution . . . 62
6.7.7 Student's t-Distribution . . . 63
6.7.8 Double Exponential (Laplace) Distribution . . . 63

7 Jointly Distributed Random Variables 64
7.0.1 The Jointly Discrete Case . . . 64
7.0.2 The Jointly Continuous Case . . . 65
7.0.3 The Mixed Case . . . 66
7.1 Marginal Distributions . . . 66
7.2 Conditional Distributions . . . 67
7.3 Expected Value . . . 69
7.3.1 Law of Total Expectation . . . 70
7.4 Variance . . . 73
7.4.1 Law of Total Variance . . . 74
7.5 Covariance . . . 76
7.5.1 Correlation . . . 77

8 Random Variable Transformations 79
8.1 Method of CDFs . . . 79
8.2 Method of Convolutions . . . 83
8.2.1 Convolutions . . . 83
8.2.2 Within Probability . . . 83
8.3 Method of Jacobians . . . 84
8.4 Method of Moment Generating Functions . . . 90

9 Order Statistics 91

10 Cheat Sheet 95
10.1 Probability Functions . . . 95
10.2 Expected Values, Variances, and Moment Generating Functions . . . 96

11 Afterword **98**

1 Foreword

I can't take credit for anything written here. Everything was transcribed from either Dr. Torcaso's lecture notes or the notes of the teaching assistants for the class. That being said, I can take credit for the anecdotes I added sometimes.

I tried to be consistent with notation, e.g. using $\mathcal{U}$ to denote a continuous uniform distribution versus just U, but it's inevitable that I missed it somewhere. I guess just use context clues in that case. Also, I tried to make sure that I didn't talk about a concept before using it in an example, e.g. iid random variables or covariance, but I'm also pretty sure I did do that. If this document was actually being looked at or mattered whatsoever I might be inclined to go fix that, but it isn't, and I'm not going to.

I honestly hope that Dr. Torcaso never dies and keeps lecturing forever because he's a wonderful teacher and not only has a love for math but also an amazing talent for teaching.

2 Basics of Probability

2.1 Kolmogorov Axioms of Probability

2.2 Sample Spaces

2.3 DeMorgan's Laws

2.4 Boole Inequalities

2.5 Conditional Probabilities

2.6 Law of Total Probability

2.7 Bayes' Theorem

2.8 Independence of Events

3 Basics of Counting

3.1 Basic Counting Principle

3.2 Bijective Counting

3.3 Partitioning Sets

3.4 Partitions of an Integer

4 Random Variables

4.1 Discrete Random Variables

A discrete random variable is a random variable which may take on only a countable number of distinct values. The probability of discrete random variables can be described using probability mass functions. We can define a random variable to be discrete provided its image is a discrete set. Using set notation, the random variable X is considered discrete if

$$\{x : X(\omega) = x \text{ for some } \omega \in \Omega\}$$

is a finite or countable infinite set, k is an outcome of some experiment, and ω is a sample point in the sample space Ω.

4.1.1 Probability Mass Function

A probability mass function (PMF) is a real-valued function defined on the sample space of an experiment which gives the probability of a discrete random variable. A PMF is different from a probability density function (PDF) in that the latter is associated with continuous rather than discrete random variables.

PMFs require the following to be true: For a random variable X and sample points $x_1, x_2, \ldots, x_n$ with probability measure P,

- $P(X = x_i) \geq 0$
- $\sum_{i=1}^{n} P(X = x_i) = 1$

i.e., they must follow the Kolmogorov Axioms of nonnegativity and normalization.

Consider the following example that illustrates how we would go about showing a function is a PMF and satisfies both of these conditions: The Riemann-Zeta function,

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

is a very important function in number theory. The zeta distribution, denoted Zeta(s), with parameter $s > 1$ is a random variable X with the PMF given by

$$P(X = x) = \frac{k^{-s}}{\zeta(s)} \quad \text{for } x \geq 1$$

To verify that Zeta(s) is a PMF, we have to show that for all $x \geq 1$, Zeta(s) ≥ 1 and that the sum of all the probabilities of x equals 1. First, nonnegativity: Since $x \geq 1$ clearly $x^s \geq 1$, for all $s > 1$ and therefore also $x^{-s} \geq 1$ for all $s > 1$. Now, normalization:

$$\begin{aligned}
\sum_{x=1}^{\infty} P(X = x) &= \sum_{x=1}^{\infty} \frac{x^{-s}}{\zeta(s)} \\
&= \frac{1}{\zeta(s)} \sum_{x=1}^{\infty} \frac{1}{x^s}^{\dagger} \\
&= \frac{\zeta(s)}{\zeta(s)} = 1
\end{aligned}$$

†This is exactly the Riemann-Zeta function as defined above!
Thus, we've shown that Zeta(s) is a PMF.

4.1.2 Cumulative Distribution

The cumulative distribution function (CDF) of a real-valued random variable X, evaluated at x, is the probability that X will take any value less than or equal to x. The functional form of the CDF is given by

$$\begin{aligned} F : \mathbb{R} \to [0,1];\ X \mapsto P(X \leq x) \\ \iff F(X) = P(X \leq x) \end{aligned}$$

The CDF always has the following properties:

- If $X < Y$, then $F(X) \leq F(Y)$ i.e., the CDF is a non-decreasing function (this fact stems from the monotonicity of probability).
- F is a right-continuous function (the limit from the right always *equals* the value of the function) that always possesses left limits.
- The limit of $F(X)$ as $X \to -\infty$ is 0 and the limit of $F(X)$ as $X \to \infty$ is 1.

If we know the CDF of a discrete random variable, we can recover the PMF of that variable using the following equations:

$$\begin{aligned} P(a < k \leq b) &= F(b) - F(a) \\ P(a \leq k \leq b) &= F(b) - F(a-) \\ P(a < k < b) &= F(b-) - F(a) \\ P(a \leq k < b) &= F(b-) - F(a-) \end{aligned}$$

Here, Dr. Torcaso's notation of $F(a-) = \lim_{x \to a^-} F(X)$ is adopted for convenience.

4.2 Continuous Random Variables

Continuous random variables have cumulative distribution function which are continuous everywhere on the real line. Because of this, at any single point the random variable has a probability of zero. Thus, we assign positive probabilities to *intervals* on the probability distribution.

4.2.1 Probability Density Function

The probability density function (PDF) of random variable is a function whose value at any given sample point can be interpreted as providing a *relative likelihood* that the value of the random variable would equal that sample point. The notion of a probability *density* as opposed to a probability *mass* that we saw with discrete random variables arises from the following issue: Recall that the CDF of a discrete random variable is given by:

$$F_X(x) = P(X \leq x)$$

Now, let $x \in \mathbb{R}$. For a positive $h \in \mathbb{R}$,

$$\begin{aligned} P(x - h < X \leq x) &= P(X \leq h) - P(X \leq x - h) \\ &= F_X(x) - F_X(x - h) \end{aligned}$$

Since F_X is continuous on $\mathbb{R}$ and if we take the limit as $h \to 0^+$,

$$P(x - h < X \leq x) = P(X = x) = 0$$

Instead, if we consider not the probability mass at x but instead the probability at x, namely

$$\frac{P(x-h < X \leq x)}{h} = \frac{F_X(x) - F_X(x-h)}{h}$$

which affords the probability *density* near x. Now, when taking the limit as $h \to 0^+$, assume that this expression exists and denote it $F'_X(x)$. Then, $f_X(x)$ the PDF, is given by

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$$

$$f_X(x) = \frac{d}{dx} F_X(x)$$

Taking the limit as $x \to \infty$,

$$\int_{-\infty}^{\infty} f_X(t)dt = 1$$

which satisfies our intuition for the total probability of a probability distribution being 1.

More generally, in order for a function f to be a PDF, it must satisfy the following requirements:

1. $f : \mathbb{R} \to \mathbb{R}, f(x) \geq 0 \ \ \forall\, x \in \mathbb{R}$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

If a random variable X has a PDF $f_X(x)$ then we compute the probability mass of X via

$$P(a < X < b) = \int_a^b f_X(x)dx$$

Showing a Function is a PDF

Consider the following function:

$$f(x) = \begin{cases} \frac{3}{8}x^2 & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

To show that this function satisfies the requirements for a PDF we have to show that it is clearly defined on the entire real line and that integrating the function from $-\infty \to \infty$ evaluates to 1: Clearly, this function is well defined on the entire real line. Now, let's integrate the function from $-\infty \to \infty$:

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)dx &= \int_{-\infty}^{0} 0\, dx + \int_0^2 \frac{3}{8}x^2\, dx + \int_2^{\infty} 0\, dx \\ &= \int_0^2 \frac{3}{8}x^2\, dx \\ &= \left[\frac{1}{8}x^3\right]_0^2 \\ &= 1 \end{aligned}$$

Thus, f is a PDF. This example also provides a good example of the *support* of a function.

Support (Essential Domain)

The *support* of a function, also referred to as the *essential domain*, is the subset of the domain containing all the elements of a function which are not mapped to zero. In set-notation, we define the support of f as:

$$\text{supp}(f) = \{x \in X : f(x) \neq 0\}$$

In the above example, the support of f is the interval $[0, 2]$. The notation of the essential domain helps us simplify computing the integral of a probability distribution because it restricts the interval of the integral.

4.2.2 Cumulative Distribution

The cumulative distribution function (CDF) of a real-valued random variable X, evaluated at x, is the probability that X will take any value less than or equal to x. For continuous random variables, the CDF can be expressed as the integral of its respective PDF:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$$

By the fundamental theorem of calculus, then, we have that:

$$F'_X(x) = f_X(x)$$

This relationship will offer an easier way of calculating the PDF of a continuous random variable.

4.3 Independent and Identically Distributed Random Variables

Independent and identically distributed random variables, oftentimes referred to as “iid" random variables is a set of variables which are (as the name implies) independent and having the same probability distribution. Oftentimes, this set is written as

$$X_1, X_2, \ldots, X_n \sim \text{iid} f$$

where f is the probability distribution of any of the X_i's. iid random variables are useful when considering order statistics and some variable transforms, as well as the central limit theorem.

4.4 Expected Value

I used to have the expected value as a subsection of discrete and continuous random variables. However, after learning more about it I've realized that it deserves its own section. By itself, the expected value can be thought of as the mean of a random variable or distribution. For this reason, the symbol μ is often used to denote the expected value.

The expected value for some discrete random variable X is given by

$$E(X) = \sum_X x \cdot P(X = x)$$

where X takes only finitely many values or $\sum_X |x| P(X = x) < \infty$. Otherwise, the expected value does not exist.

The expected value of a continuous random variable is given by

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

where $f_X(x)$ is the PDF of X. Note that this expectation will exist *only* when $\int_{-\infty}^{\infty} |x| f_X(x)\, dx < \infty$ i.e., the integral converges.

$E(X)$ is sometimes referred to as the first moment of X. In a similar vein, $E(X^2)$ is referred to as the second moment of X, $E(X^3)$ is referred to as the third moment of X, and so on. A sample calculation of the expected value for a discrete random variable is illustrated below:

x	$P(X = x)$
-2	0.1
-1	0.1
0	0.1
1	0.3
2	0.2
4	0.2

Consider the PMF above. The expected value of this PMF is given by

$$\begin{aligned}
E(X) &= \sum_X x \cdot P(X = x) \\
&= -2P(X = -2) - 1P(X = -1) + 0P(X = 0) + 1P(X = 1) + 2P(X = 2) + 4P(X = 4) \\
&= -2(0.1) - 1(0.1) + 0(0.1) + 1(0.3) + 2(0.2) + 4(0.2) \\
&= 1.2
\end{aligned}$$

Similarly, we can find the expected value of this random variable X *squared*, X^2:

$$\begin{aligned}
E(X^2) &= \sum_X x^2 \cdot P(X^2 = x) \\
&= 0P(X^2 = 0) + 1P(X^2 = 1) + 4P(X^2 = 4) + 16P(X^2 = 16) \\
&= 0P(X = 0) + 1[P(X = 1) + P(X = -1)] + 4P[P(X = 2) + P(X = -2)] \\
&\qquad + 16[P(X = 4) + P(X = -4)] \\
&= 0[0.1] + 1[0.3 + 0.1] + 4[0.2 + 0.1] + 16[0.2 + 0] \\
&= 0 + .4 + 1.2 + 3.2 \\
&= 4.8
\end{aligned}$$

Now let's try an example involving a continuous random variable. Compute the expected value of the following PDF:

$$f(x) = \begin{cases} \frac{3}{8}x^2 & 0 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

Okay!

$$\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_0^2 x\left(\frac{3}{8}x^2\right) dx \\
&= \frac{3}{8}\int_0^2 x^3\, dx = \frac{3}{8}\left[\frac{1}{4}x^4\right]_0^2 \\
&= \frac{3}{8}(4) = \frac{3}{2}
\end{aligned}$$

Note that we utilized the support of f in order to simplify the integral.

Here is a quick example illustrating how we might use some iid random variables to simplify solving an expected value question: Suppose we are randomly dealt 13 cards from a standard deck of 52 and let X be the random variable which counts the number of suits in our hand. Compute the expected value of the number of suits in our hand, $E(X)$. To do this, we'll let $X = X_1 + X_2 + X_3 + X_4$ where X_i represents a unique suit in the deck and

$$X_i = \begin{cases} 1 & \text{if suit } i \text{ is in the hand} \\ 0 & \text{otherwise} \end{cases}$$

for suits $i = 1, 2, 3, 4$. By the linearity of the expectation, we have that

$$E(X) = E(X_1 + X_2 + X_3 + X_4) = E(X_1) + E(X_2) + E(X_3) + E(X_4) = 4E(X_1)$$

since all X_i have the same distribution. Now we can use the definition of the expected value to compute the expectation of X:

$$\begin{aligned} E(X) = 4E(X_1) &= 4[1 \cdot P(X_1 = 1) + 0 \cdot P(X_1 = 0)] \\ &= 4[1 - P(X_1 = 0)] \\ &= 4\left[1 - \frac{\binom{39}{13}}{\binom{52}{13}}\right] \approx 3.9488 \end{aligned}$$

4.4.1 Law of the Unconscious Statistician (LOTUS)

The law of the unconscious statistician (LOTUS) says that for any random variable X and a "function" $g : \mathbb{R} \to \mathbb{R}$, then

$$\begin{aligned} E(g(x)) &= \sum_X g(x)P(X = x) \quad \text{if } X \text{ is discrete} \\ &= \int_{-\infty}^{\infty} g(x) f_X(x)\, dx \quad \text{if } X \text{ is continuous} \end{aligned}$$

as long as the expected value exists. Consider some consequences of LOTUS: Take, for example, the function $g(t) = at + b$ where $a, b \in \mathbb{R}$ are fixed constants. Then,

$$E(g(X)) = aE(X) + b$$

To show why this works, we can consider the definition of the expected value. First, for the discrete case:

$$\begin{aligned} E(g(X)) &= \sum_X g(X)P(X = x) \\ &= \sum_X (ax + b)P(X = x) \\ &= \sum_X (axP(X = x) + bP(X = x)) \\ &= a\sum_X xP(X = x) + b\sum_X P(X = x) \\ &= a(E(X)) + b(1) \\ &= aE(X) + b \end{aligned}$$

Now, for the continuous case:

$$\begin{aligned}
E(g(X)) &= \int_{-\infty}^{\infty} g(x) f_X(x) \\
&= \int_{-\infty}^{\infty} (ax + b) f_X(x) \\
&= \int_{-\infty}^{\infty} (ax f_X(x) + b f_X(x)) \\
&= a \int_{-\infty}^{\infty} x f_X(x)) + b \int_{-\infty}^{\infty} f_X(x) \\
&= a(E(X)) + b(1) \\
&= aE(X) + b
\end{aligned}$$

From this derivation, we find that the expected value of a constant is just that constant, e.g. $E(b) = b$. This could also be thought of as taking the constant out i.e., $E(b) = bE(1)$, and finding the expected value of 1, which is 1.

The LOTUS us helpful when trying to evaluate the expectation of some expression in a random variable. Consider the case of a discrete random variable. Suppose $X \sim \text{Poisson}(\lambda)$ and we are interested in $E(\frac{1}{1+X})$. We can evaluate the expectation in the following way:

$$\begin{aligned}
E\left(\frac{1}{1+X}\right) &= \sum_X \left(\frac{1}{1+x}\right) P(X = x) \\
&= \sum_{x=0}^{\infty} \left(\frac{1}{1+x}\right) \frac{e^{-\lambda}\lambda^x}{x!} \\
&= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{(x+1)!} \\
&= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^{x+1}}{\lambda(x+1)!} \\
&= \frac{e^{-\lambda}}{\lambda} \sum_{x=0}^{\infty} \frac{\lambda^{x+1}}{(x+1)!}
\end{aligned}$$

Now, we'll make the substitution $j = x + 1$:

$$\begin{aligned}
&= \frac{e^{-\lambda}}{\lambda} \sum_{j=1}^{\infty} \frac{\lambda^j}{j!} \\
&= \frac{e^{-\lambda}}{\lambda} \left[\sum_{j=0}^{\infty} \frac{\lambda^j}{j!} - \frac{\lambda^0}{0!} \right] \\
&= \frac{e^{-\lambda}}{\lambda} \left[e^{\lambda} - 1 \right] \\
&= \frac{1 - e^{-\lambda}}{\lambda}
\end{aligned}$$

4.4.2 Linearity of the Expected Value

One fact we semi-took advantage of with LOTUS was the linearity of the expected value of a random variable. This property of the expected value is *unparalleled* by all others. Formally, for a

random variable X,

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

What's amazing is that this property requires nothing more than the fact that the expectations of the random variables being summed exist.

A concrete example of this fact is the relationship between the expected value of the binomial distribution and that of n independent Bernoulli trials (both of these distributions are covered in more detail later): If we let $Y \sim \text{binomial}(n, p)$ and $X \sim \text{Bernoulli}(p)$, and we already know that $E(X) = p$, then

$$\begin{aligned} E(Y) &= E(X_1) + E(X_2) + E(X_3) + \cdots + E(X_n) \\ &= p + p + p + \cdots + p \\ &= np \end{aligned}$$

which agrees with our formulation for the expected value of a binomial distribution.

Another good example of this fact is the relationship between the geometric and hypergeometric distributions. If we let $X \sim \text{hypergeometric}(n, M, N)$ then we can express X as the sum of n independent random variables each with a Bernoulli distribution. That is,

$$X = \sum_{i=1}^{n} Y_i \quad \text{where } Y_i \sim \text{Bernoulli}\left(\frac{M}{N}\right)$$

Thus, the expected value of X becomes

$$\begin{aligned} E(X) = E\left(\sum_{i=1}^{n} Y_i\right) &= E(Y_1 + Y_2 + \cdots + Y_n) \\ &= E(Y_1) + E(Y_2) + \cdots + E(Y_n) \\ &= \frac{M}{N} + \frac{M}{N} + \cdots + \frac{M}{N} \\ &= \frac{nM}{N} \end{aligned}$$

which agrees with the known result for the expectation of the hypergeometric distribution.

A useful application of this linearity is illustrated in the following example: Suppose $X_1, X_2, \ldots, X_n$ are random variables each with mean μ and variance σ^2. In statistics,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

where $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is the sample mean, is a measure used to estimate the variance of the population from which the sample is drawn. Given that $E(\overline{X}) = \mu$ and $\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$, show that $E(s^2) = \sigma^2$.

Here, we'll use the linearity of the expectation and express $E(s^2)$ in a more friendly manner:

$$E(s^2) = E\left[\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2\right] = \frac{1}{n-1} E\left[\sum_{i=1}^{n} (X_i - \overline{X})^2\right]$$

Before trying to evaluate that, let's first play with the summation inside of the expectation:

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \overline{X})^2 &= \sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X} + \overline{X}^2) \\ &= \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} 2X_i\overline{X} + \sum_{i=1}^{n}\overline{X}^2 \\ &= \sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i + n\overline{X}^2\end{aligned}$$

Recall that we are given $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \iff n\overline{X} = \sum_{i=1}^{n} X_i$. Thus,

$$\begin{aligned}&= \sum_{i=1}^{n} X_i^2 - 2\overline{X}(n\overline{X}) + n\overline{X}^2 \\ &= \sum_{i=1}^{n} X_i^2 - n\overline{X}^2\end{aligned}$$

Now, we can rewrite $E(s^2)$ in the following way:

$$E(s^2) = \frac{1}{n-1}E\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right]$$

From the linearity of the expectation, we can rewrite this as

$$E(s^2) = \frac{1}{n-1}\left[E\left(\sum_{i=1}^{n} X_i^2\right) - nE\left(\overline{X}^2\right)\right]$$

Additionally, due to the linearity of summation, we can express this as

$$E(s^2) = \frac{1}{n-1}\left[\sum_{i=1}^{n} E(X_i^2) - nE(\overline{X}^2)\right]$$

From here, there are a few things we could try and do to simplify this calculation. The first thought that comes to mind might just be evaluating the second moment of each of those random variables and simplifying the expression. Being the clever mathematicians we are, however, we can use more of the information given to us in the problem to find these expectations. Using the definition of the variance, $\text{Var}(X) = E(X^2) - [E(X)]^2$, we can find $E(X_i^2)$ and $E(\overline{X}^2)$:

$$\begin{aligned}\text{Var}(X_i) &= E(X_i^2) - [E(X_i)]^2 \\ \sigma^2 &= E(X_i^2) - \mu^2 \iff E(X_i^2) = \sigma^2 + \mu^2\end{aligned}$$

Similarly,

$$\begin{aligned}\text{Var}(\overline{X}) &= E(\overline{X}^2) - [E(\overline{X})]^2 \\ \frac{\sigma^2}{n} &= E(\overline{X}^2) - \mu^2 \iff E(\overline{X_i}^2) = \frac{\sigma^2}{n} + \mu^2\end{aligned}$$

Now, plug these back into our expression for the expectation of s^2:

$$\begin{aligned}
E(s^2) &= \frac{1}{n-1}\left[\sum_{i=1}^{n}(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right] \\
&= \frac{1}{n-1}\left[n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right] \\
&= \frac{1}{n-1}\left[n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2\right] \\
&= \frac{1}{n-1}\left[\sigma^2(n-1)\right] \\
&= \sigma^2
\end{aligned}$$

In the context of the problem, this results makes perfect sense. If s^2 is used to estimate the variance from a population we'd expect that the expected value (or mean) of s^2 to be σ^2, the actual variance of the population. Otherwise, s^2 would be a pretty poor way to estimate the variance.

The linearity of the expected value also holds for products of random variables *when those random variables are independent*. Formally, this equality can be stated as follows: For any random variables $X_1, \ldots, X_n$ and functions $g_1, \ldots, g_n : \mathbb{R} \to \mathbb{R}$, then:

$$E\left[\prod_{i=1}^{n} g_i(X_i)\right] = \prod_{i=1}^{n} E[g_i(X_i)]$$

As an example, for the case of two random variables X and Y we have that

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

4.4.3 When the Expected Value Does Not Exist

An interesting case of the expected value not existing for a discrete random variable takes advantage of a consequence of one of our favorite functions from complex analysis: The Riemann-Zeta function says that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

Thus, we can create a PMF out of this equation:

$$P(X = n) = \frac{6}{\pi^2 n^2} \quad \text{for } n = 1, 2, 3, \ldots$$

Now, let's try a calculate the expected value of this PMF:

$$\begin{aligned}
E(X) &= \sum_{n=1}^{\infty} n \cdot P(X = n) \\
&= \sum_{n=1}^{\infty} n \cdot \frac{6}{\pi^2 n^2} \\
&= \frac{6}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n}
\end{aligned}$$

We can show that this series diverges using some calculus techniques, but the proof is trivial and left as an exercise for the reader. Thus, there is no expected value for this PMF.

4.4.4 Expectation When the Random Variable is Nonnegative

An interesting and useful way to compute the expectation of a continuous random variable when the PDF is strictly nonnegative is to use the complement of the CDF. Consider the following: If a random variable X with CDF given by F_X is strictly nonnegative on all of $\mathbb{R}$, then

$$E(X) = \int_0^\infty P(X > t)\, dt = \int_0^\infty [1 - F_X(t)]\, dt$$

is the expected value of X. To show why this is true, consider the definition of the expected value:

$$\begin{aligned} E(X) &= \int_0^\infty x f_X(x)\, dx \\ &= \int_0^\infty \int_0^x f_X(x)\, du dx \\ &= \int_0^\infty \left[\int_u^\infty f_X(x)\, dx \right] du \\ &= \int_0^\infty \left[1 - \int_0^u f_X(x)\, dx \right] du \\ &= \int_0^\infty \left[1 - F_X(u) \right] du \end{aligned}$$

And we've shown what we sought to prove. Note that our integration begins at 0 in the first line because the random variable is defined to be nonnegative and therefore has a nonnegative support.

4.5 Variance

The variance is the expectation of the squared deviation of random variable from its mean. Symbolically,

$$\text{Var}(X) = E[(X - \mu)^2]$$

A more computationally friendly method of computing the variance is

$$\text{Var}(X) = E(X^2) - [E(X)]^2$$

which can be derived directly from the definition above (and is done below).

The variance is a measure of how spread out a data set is from its mean. Another common measure of this spread is the standard deviation, σ, which is just the square root of the variance. We can express the variance in a way that is more computationally friendly using the LOTUS: Letting $g(X) = (X - \mu)^2$, then

$$\begin{aligned} E(g(X)) &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - E(2\mu X) + E(\mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2 \end{aligned}$$

As an example, consider the following PMF:

x	$P(X = x)$
-2	0.1
-1	0.1
0	0.1
1	0.3
2	0.2
4	0.2

We've already calculated the first and second moment of the random variable X, so all we have to do is plug in to find the variance:

$$\text{Var}(X) = \sigma^2 = E(X^2) - \mu^2 = 4.8 - 1.2 = 3.6$$

4.5.1 "LOTUS" for the Variance

Just as we saw how the expectation of a random variable changed under some affine transformation, a similar result can be generalized for the variance. For a function $g(t) = at + b$ where $a, b \in \mathbb{R}$ are fixed constants and a random variable X,

$$\text{Var}(g(X)) = a^2\text{Var}(X)$$

when the expectation of X exists and is finite.

4.5.2 Nonnegativity of the Variance

Here, we'll start with a remark: For any random variable X, $(X - \mu^2) > 0$, and therefore

$$E((X - \mu)^2) = \sum_X (k - \mu)^2 P(X = x)$$

Thus, the variance is always nonnegative. Moreover,

$$\begin{aligned} E((X - \mu)^2) &\geq 0 \\ E(X^2) - [E(X)]^2 &\geq 0 \\ \iff E(X^2) &\geq [E(X)]^2 \\ \sqrt{E(X^2)} &\geq E(X) \end{aligned}$$

Thus, the magnitude of the mean of a random variable is always less than the square root of the second moment of the same random variable. This relationship is known as the Lyapunov inequality.

4.5.3 Final Remarks

A final remark we will make here regards the importance of a random variable's mean and variance. For a random variable X, if we know the mean $\mu = E(X)$ and variance $\sigma^2 = \text{Var}(X)$, then we claim that

$$P(|X - \mu| \geq x\sigma) \leq \frac{1}{x^2}$$

This relationship is known as Chebyshev's inequality and it says that no more than a certain fraction of values can be more than a certain distance from the mean of the random variable. Specifically, no more than $1/x^2$ of the distribution's values can be more than x standard deviations away from the mean. This inequality is related to the law of large numbers.

4.6 Central Limit Theorem

The Central Limit Theorem (CLT) states that if we have independent and identically distributed random variables $X_1, X_2, \dots, X_n$ with finite mean and finite variance μ and σ^2 respectively. Let $S_n = X_1 + X_2 + \cdots + X_n$. Then, for any $-\infty < a < b < \infty$ we have:

$$\bigstar \lim_{n\to\infty} \left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b \right) = \Phi(b) - \Phi(a) = \int_a^b \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}}\, dz \bigstar$$

Loosely speaking, what this random variable says is that as $n \to \infty$

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \to Z$$

where $n\mu$ is the mean of S_n and $\sigma\sqrt{n}$ is the standard deviation of S_n. Consequently, for large, fixed values of n, $S_n \approx n\mu + \sigma\sqrt{n}Z \sim \mathcal{N}(n\mu, \sigma^2 n)$.

The CLT is indeed a limit result. When n is *large enough* we can use the CDF of the standard normal distribution to estimate the CDF of the sum of iid random variables. This begs the question, how large must n be to yield a decent approximation? Below are some loose guidelines:

- When the population distribution (i.e. identical distribution) is symmetric about its mean, then n does not need to be that large. Oftentimes, $n \geq 20$ is good enough and in some special cases $n = 3$ may even by used.
- When the population distribution is only “slightly skewed" or nearly symmetric, then $n \geq 30$ may be considered “large enough" to yield a decent approximation.
- Generally, the more skewed a distribution, the larger n must be to provide a good estimation

Let’s consider a simple example to illustrate how we could estimate a distribution using the CLT: Suppose we roll a 6-sided die 3 times and let S be the sum of the rolls. Compute $P(9 \leq S \leq 12)$. The exact answer to this question is $\frac{13}{27} \approx 0.4815$. Now, let’s try and estimate this probability using the CLT (note that we know before performing this calculation that this will be a bad estimation because $n = 3$ is very small):

$$\begin{aligned}
\mu = E(S) &= E\left(\sum_{i=1}^{6} X_i \right) = \sum_{i=1}^{6} E(X_i) = \sum_{i=1}^{6} x_i P(X_i = x_i) \\
&= \sum_{i=1}^{6} \frac{1}{6} x_i = \frac{7}{2} \\
\sigma^2 = \mathrm{Var}(S) &= E(S^2) - [E(S)]^2 = \sum_{i=1}^{6} E(X_i^2) - \left(\frac{7}{2}\right)^2 = \sum_{i=1}^{6} x_i^2 P(X_i^2 = x_i) - \frac{49}{4} \\
&= \sum_{i=1}^{6} \frac{1}{6} x_i^2 - \frac{49}{4} = \frac{1}{6}\left(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2\right) - \frac{49}{4} \\
&= \frac{35}{12}
\end{aligned}$$

We have the mean and variance of the distribution of S, so we can apply the CLT:

$$P(9 \le S \le 12) = P\left(\frac{9 - 3(\frac{7}{2})}{\sqrt{3(\frac{35}{12})}} \le \frac{S - n\mu}{\sigma\sqrt{n}} \le \frac{12 - 3(\frac{7}{2})}{\sqrt{3(\frac{35}{12})}}\right) = P\left(-0.5071 \le \frac{S - n\mu}{\sigma\sqrt{n}} \le 0.5071\right)$$
$$\approx P(-0.5071 \le Z \le 0.5071)$$
$$= \Phi(0.5071) - \Phi(-0.5071)$$
$$= 0.6950 - 0.3050 = 0.3900$$

This is our approximation for the probability using the CLT. Obviously, it sucks. Why? Well, recall that our population size was $n = 3$ which is nowhere near large enough to provide an accurate estimation using the CLT. However, there is a way to make our estimation better, using the *continuity correction* (discussed later).

Since $n = 3$ was too small, for what value of n can we be sure that the normal distribution will afford an accurate estimate for the binomial? Different sources will cite different number. One common rule of thumb is that $np \ge 5$ and $n(1 - p) \ge 5$ must be true to afford an accurate estimate. Another guideline, provided by Dr. Torcaso, is that $np(1 - p) \ge 5$ must hold to afford a good estimate. Using this rule, let's compute the value of n necessary for an accurate approximation for different probabilities p:

p	$n \ge$
0.5	20
0.2	32
0.1	56
0.01	560

4.6.1 The Continuity Correction

In general, the *continuity correction* is an adjustment that is made when a discrete distribution is approximated by a continuous distribution. In our case, we'll be interested in applying the continuity correction when estimating integer-valued distributions with the standard normal distribution using the central limit theorem. For this correction we can make the substitutions:

$$P(a \le X \le b) = P\left(a - \frac{1}{2} \le X \le b + \frac{1}{2}\right) \quad \& \quad P(a < X < b) = P\left(a + \frac{1}{2} \le X \le b - \frac{1}{2}\right)$$

Consider the previously mentioned example which involved estimating the binomial distribution using a standard normal distribution. Here, we claim that

$$P(9 \le S \le 12) = P(8.5 \le S \le 12.5)$$

due to that fact that the binomial distribution is integer-valued. So, let's try estimating the distribution using a standard normal again, this time using the new bounds:

$$P(8.5 \le S \le 12.5) = P\left(\frac{8.5 - 3(\frac{7}{2})}{\sqrt{3(\frac{35}{12})}} \le \frac{S - n\mu}{\sigma\sqrt{n}} \le \frac{12.5 - 3(\frac{7}{2})}{\sqrt{3(\frac{35}{12})}}\right) = P\left(-0.6761 \le \frac{S - n\mu}{\sigma\sqrt{n}} \le 0.6761\right)$$
$$\approx P(-0.6761 \le Z \le 0.6761)$$
$$= \Phi(0.6761) - \Phi(-0.6761)$$
$$= 0.7505 - 0.2495 = 0.5010$$

A-ha! This is a far better approximation for the binomial distribution provided. Recall that the actually probability was ≈ 0.4815.

Here's another example where we can employ the continuity correction: Suppose we toss a fair coin 100 times and let S count the number of heads we see. Estimate $P(35 \leq S \leq 65)$ (the actual probability is something like ≈ 0.9982). First, we'll evaluate the mean and variance of S:

$$\mu = E(S) = np = 100\left(\frac{1}{2}\right) = 50$$
$$\sigma^2 = \text{Var}(S) = np(1-p) = 100\left(\frac{1}{2}\right)\left(1-\frac{1}{2}\right) = 25$$

Cool. Now let's estimate the distribution. Note that I'll apply the continuity correction on the bounds because we'll be estimating a binomial distribution with a standard normal distribution (again):

$$\begin{aligned} P(34.5 \leq S \leq 65.5) = P\left(\frac{34.5-50}{\sqrt{25}} \leq \frac{S-n\mu}{\sigma\sqrt{n}} \leq \frac{65.5-50}{\sqrt{25}}\right) &= P\left(-3.1 \leq \frac{S-n\mu}{\sigma\sqrt{n}} \leq 3.1\right) \\ &\approx P(-3.1 \leq Z \leq 3.1) \\ &= \Phi(3.1) - \Phi(-3.1) \\ &= 0.9990 - 0.0009 = 0.9981 \end{aligned}$$

An excellent estimation! The fact that this distribution was symmetric about its mean (because $p = 0.5$) and we had a very large n resulted in this accuracy. Note that without using the continuity correction our estimated probability would have been ≈ 0.9973 which is also very accurate. Thus, with such a large population size, the continuity correction is not always absolutely necessary.

4.7 Moment Generating Functions

The moment generating function (MGF) of a random variable is an alternative specification of its probability distribution. As the name implies, the MGF of a probability distribution is used to compute the moments of the distribution. In particular, the nth moment of a random variable is the nth derivative of the random variables MGF evaluated at 0. Unlike the characteristic function, the MGF of a random variable does not always exist.

The MGF of a random variable X is given by

$$M(\theta) = E(e^{\theta X}), \quad \theta \in \mathbb{R}$$

wherever the expectation exists. The MGF of a random variable can be computed by brute force using the Law of the Unconscious Statistician,

$$E(e^{\theta X}) = \int_{-\infty}^{\infty} e^{\theta x} f(x)\, dx$$

where $f(x)$ is the PDF of the random variable. Here are a few helpful properties of the moment generating function:

1. If $M_X(\theta)$ exists, then $E(X^k)$ exists and is finite for all $k > 0$.
2. If $M_X(\theta)$ and $M_Y(\theta)$ exist and $M_X(\theta) = M_Y(\theta)$ on an open interval containing θ, then for all θ where the MGFs are defined $M_X(\theta) = M_Y(\theta)$ and X and Y have the same probability distribution. That is, MGFs uniquely determine a probability distribution.

3. If $M_X(\theta)$ exists then

$$M_X(\theta) = E(e^{\theta X}) = E\left(\sum_{k=1}^{\infty} \frac{(\theta x)^k}{k!}\right) = \sum_{k=0}^{\infty} \frac{\theta^k}{k!} E(X^k)$$
$$= 1 + \theta E(X) + \frac{\theta^2}{2!} E(X^2) + \frac{\theta^3}{3!} E(X^3) + \dots$$

4. The kth derivative with respect to θ of $M_X(\theta)$ evaluated at $\theta = 0$ is the kth moment of X, $E(X^k)$ (Hence the name, "moment generating function").

5. If $X_1, X_2, \dots, X_n$ are independent random variables possessing respecting MGFs $M_1, M_2, \dots, M_n$ then the MGF of $S_n = X_1 + \cdots + X_n$ is given by

$$M_{S_n}(\theta) = \prod_{i=1}^{n} M_i(\theta)$$

and if the random variables X_i are identically distributed,

$$M_{S_n}(\theta) = [M(\theta)]^n$$

6. If X has the MGF $M(\theta)$ then for all $a, b \in \mathbb{R}$, $a + bX$ has the MGF $e^{a\theta} M(b\theta)$.

Another interesting aspect of the MGF is its relation to the Laplace transform. Recall that the Laplace transform is given by

$$\mathcal{L}\{f(s)\} = \int_0^{\infty} e^{-st} f(t)\, dt \quad \text{for } t > 0$$

where s is any complex number. We define the *two-sided* or *bilateral Laplace transform* by extending the limits of integration to cover the entire real line:

$$\mathcal{B}\{f(s)\} = \int_{-\infty}^{\infty} e^{-st} f(t)\, dt$$

By replacing s with $-\theta$ in the bilateral Laplace transform, we arrive at the MGF of a random variable:

$$\mathcal{B}\{f(-\theta)\} = \int_{-\infty}^{\infty} e^{-\theta t} f(t)\, dt$$

Let's shift gears a little bit to recall an important result from calculus: For some constant $c \in \mathbb{R}$,

$$\lim_{n\to\infty} \left(1 + \frac{c}{n}\right) = e^c$$

is a known limit for e. Moreover,

$$\lim_{n\to\infty} \left(1 + \frac{c}{n} + \frac{2}{n^2}\right) = e^c$$
$$\lim_{n\to\infty} \left(1 + \frac{c}{n} + \frac{69}{n^{69}}\right) = e^c$$
$$\lim_{n\to\infty} \left(1 + \frac{c}{n} + \frac{3}{n^{3/2}} + \frac{5}{n^{5/2}} + \frac{7}{n^{7/2}}\right) = e^c$$

are *also* all limits for e. In general, we can say that this limit converges to e^c *whenever* a function is added which goes to zero faster than $1/n$ as $n \to \infty$. We can express this fact using *little-o notation*, expressed as $o(g(n))$ for a function g. By definition if a function f is $f(n) = o(g(n))$, then:

$$\lim_{n\to\infty}\left(\frac{f(n)}{g(n)}\right) = 0$$

Now, consider the following, important example (I'll call it ⋆ because I refer back to it a few times later) which highlights this uniqueness of the moment generating function: Suppose $X_1, \ldots, X_n$ are all independently and identically distributed with the MGF $M(\theta)$. Suppose further that the expected value of these random variables is $E(X_i) = \mu$ and the second moment is $E(X_i) = \sigma^2 + \mu^2$. Find the MGF of

$$Y_n = \frac{\sum_{i=1}^n (X_i - \mu)}{\sigma\sqrt{n}}$$

which represents the z-score of the sum S_n of the random variables, expressed as:

$$Y_n = \frac{S_n - E(S_n)}{\sqrt{\text{Var}(S_n)}}$$

We can find the moment generating function using the linearity of the expected value:

$$\begin{aligned}
M_{Y_n}(\theta) = E(e^{\theta Y_n}) &= E\left(\exp\left\{\theta\frac{\sum_{i=1}^n (X_i-\mu)}{\sigma\sqrt{n}}\right\}\right) = E\left(\prod_{i=1}^n e^{\frac{\theta(X_i-\mu)}{\sigma\sqrt{n}}}\right)\\
&= \prod_{i=1}^n E\left(e^{\frac{\theta}{\sigma\sqrt{n}}X_i}e^{\frac{-\theta\mu}{\sigma\sqrt{n}}}\right)\\
&= \prod_{i=1}^n e^{\frac{-\theta\mu}{\sigma\sqrt{n}}}M\left(\frac{\theta}{\sigma\sqrt{n}}\right)\\
&= \left[e^{\frac{-\theta\mu}{\sigma\sqrt{n}}}M\left(\frac{\theta}{\sigma\sqrt{n}}\right)\right]^n
\end{aligned}$$

This is the MGF of Y_n. Now, let's evaluate the term which is being raised to the power of n. To do this, recall the Maclaurin expansion for the MGF and substitute our expression within M to the series:

$$\begin{aligned}
M\left(\frac{\theta}{\sigma\sqrt{n}}\right) &= 1 + E(X)\left(\frac{\theta}{\sigma\sqrt{n}}\right) + \frac{E(X^2)}{2!}\left(\frac{\theta}{\sigma\sqrt{n}}\right)^2 + \frac{E(X^3)}{3!}\left(\frac{\theta}{\sigma\sqrt{n}}\right)^3 + \ldots\\
&= 1 + \frac{\mu\theta}{\sigma\sqrt{n}} + \frac{(\sigma^2+\mu^2)\theta^2}{2\sigma^2 n} + \frac{E(X^3)}{3!}\left(\frac{\theta^3}{\sigma^3 n^{3/2}}\right) + \ldots\\
&= 1 + \frac{\mu\theta}{\sigma\sqrt{n}} + \frac{(\sigma^2+\mu^2)\theta^2}{2\sigma^2 n} + o\left(\frac{1}{n}\right)
\end{aligned}$$

Notice that every expression beyond the third term will have a power of n greater than or equal to $\frac{3}{2}$ in the denominator in it. In other words, every term is *little-o 1 over n*. Knowing this, we can sort of "ignore" the remaining terms in the sequence. Now, we can begin to evaluate the original term which was being raised to the power of n by substituting this Maclaurin expansion in for the MGF and also using the Maclaurin expansion for e^x for the other term:

$$\left[e^{\frac{-\theta\mu}{\sigma\sqrt{n}}}M\left(\frac{\theta}{\sigma\sqrt{n}}\right)\right]^n = \left\{1 - \frac{\theta\mu}{\sigma\sqrt{n}} + \frac{\theta^2\mu^2}{2\sigma^2 n} + o\left(\frac{1}{n}\right)\right\}\left\{1 + \frac{\mu\theta}{\sigma\sqrt{n}} + \frac{(\sigma^2+\mu^2)\theta^2}{2\sigma^2 n} + o\left(\frac{1}{n}\right)\right\}$$

Expanding this, we find that many terms become little-o $1/n$:

$$\begin{aligned} &= 1 + \frac{\mu\theta}{\sigma\sqrt{n}} + \frac{(\sigma^2+\mu^2)\theta^2}{2\sigma^2 n} + o\left(\frac{1}{n}\right) \\ &\quad - \frac{\theta\mu}{\sigma\sqrt{n}} - \frac{\theta^2\mu^2}{\sigma^2 n} - o\left(\frac{1}{n}\right) \\ &\quad + \frac{\theta^2\mu^2}{2\sigma^2 n} + o\left(\frac{1}{n}\right) \\ &= 1 + \frac{\theta^2}{2n} + o\left(\frac{1}{n}\right) \end{aligned}$$

Thus, we have that

$$\left[e^{\frac{-\theta\mu}{\sigma\sqrt{n}}} M\left(\frac{\theta}{\sigma\sqrt{n}}\right)\right]^n = \left[1 + \frac{\theta^2}{2n} + o\left(\frac{1}{n}\right)\right]^n$$

and therefore

$$\lim_{n\to\infty}\left[1 + \frac{\theta^2}{2n} + o\left(\frac{1}{n}\right)\right]^n = e^{\frac{1}{2}\theta^2}$$

which is exactly the moment generating function of the standard normal distribution. This result was expected. Why?

4.7.1 Continuity Theorem of Moment Generating Functions

If we have a sequence (Y_n) of random variables that each have a moment generating function and the limit as $n \to \infty$ of $M_{Y_n}(\theta) = M_X(\theta)$ for all θ, where $M_X(\theta)$ is the MGF of a random variable X having a continuous CDF, then for all real x:

$$\lim_{n\to\infty} P(Y_n \le x) = P(X \le x)$$

In words that actually make sense, this theorem is telling us that our intuition is correct: If we have a sequence of random variables whose MGFs converge to the MGF of a random variable, then the distribution of the sequence of random variables converges to the distribution of that random variable.

4.8 Indicator Functions

The indicator function, denoted I, is a function defined on a set that indicates the membership of an element within a subset. The indicator will have a value of 1 for all elements in the subset and a value of 0 for all elements not in the set. More simply, the indicator function is a way to express nonzero parts of a function without using a piecewise definition. In set notation, the indicator function is defined as:

$$I_{[a,b]}(x) = \begin{cases} 1 & x \in [a,b] \\ 0 & x \notin [a,b] \end{cases}$$

As an example, consider once again the PDF offered in section **1.4**:

$$f(x) = \begin{cases} \frac{3}{8}x^2 & 0 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

This function can also be expressed using an indicator function:

$$f_X(x) = \frac{3}{8}x^2 I_{[0,2]}(x)$$

Intuitively, we can think of the indicator function as “flipping on" when on the support of the function. Additionally, the use of indicator functions can simplify finding the constants of integration when integrating the PDF of a random variable. If we wanted to find the expected value of f_X, instead of utilizing the support of the function we can utilize the indicator:

$$E(X) = \int_{-\infty}^{\infty} f_X(x)\,dx = \int_{-\infty}^{\infty} \frac{3}{8}x^2 I_{[0,2]}(x)\,dx = \int_0^2 \frac{3}{8}x^2\,dx$$

which affords the same value as when we didn't use indicator functions.

As an example of using indicator functions, consider the following example: Let $X \sim \text{Exp}(\lambda)$. What is the PDF of $Y = X - 3$? We can express the PDF of X as

$$f_X(x) = \lambda e^{-\lambda x} I_{[0,\infty]}(x)$$

Using the CDF method for finding the PDF of Y, we have that

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(X - 3 \leq y) = P(X \leq y + 3) \\ &= F_X(y + 3) \end{aligned}$$

Now, let's go about evaluating the CDF of X:

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} \lambda e^{-\lambda u} I_{[0,\infty]}(u)\,du \\ &= \left[-e^{-\lambda u} I_{[0,\infty]}(u) \right]_0^x \\ &= -e^{-\lambda x} I_{[0,\infty]}(x) - (-e^0 I_{[0,\infty]}(x)) = (1 - e^{-\lambda x}) I_{[0,\infty]}(x) \end{aligned}$$

Then, the CDF for $Y = X - 3$ is given by plugging in $x = y + 3$ in the CDF for X:

$$F_Y(y) = (1 - e^{-\lambda(y+3)}) I_{[0,\infty]}(y + 3) = (1 - e^{-\lambda(y+3)}) I_{[-3,\infty]}(y)$$

Now, we can find the PDF for Y using the relationship $F_Y'(y) = f_Y(y)$. The PDF for Y becomes:

$$f_Y(y) = \lambda e^{-\lambda(y+3)} I_{[-3,\infty]}(y)$$

Recall that this can be just as easily expressed using a piecewise function:

$$f_Y(y) = \begin{cases} \lambda e^{-\lambda(y+3)} & -3 \leq y \leq \infty \\ 0 & \text{otherwise} \end{cases}$$

5 Discrete Random Variables

5.1 Discrete Uniform Distribution

5.2 Bernoulli Distribution

The Bernoulli distribution, or $X \sim \text{Bernoulli}(p)$, is a discrete probability distribution of a random variable which takes the value of 1 with probability p and the value of 0 with probability $1 - p$. Informally, it can be thought of as as model for the set of possible outcomes of any single experiment that asks a yes-no question.

5.2.1 Probability Mass Function

The PMF of the Bernoulli distribution is given by:

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

In its functional form, the Bernoulli distribution is written as:

$$P(X = x) = p^x(1 - p)^{1-x} \quad \text{for } x \in \{0, 1\}$$

As we will see, the Bernoulli distribution is the least interesting of all of our discrete probability distributions. However, it is the most important as it is a building block for many other distributions.

In real life, a Bernoulli distribution might arise if we flip a coin once. In this case, $X \sim \text{Bernoulli}(0.5)$ because each side has an equal likelihood of flipping.

5.2.2 Cumulative Distribution Function

The CDF of the Bernoulli(p) distribution is given by

$$\begin{cases} 0 & \text{if } x < 0 \\ 1 - p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } 1 \geq x \end{cases}$$

5.2.3 Expected Value

The first moment of the random variable $X \sim \text{Bernoulli}(p)$ is $\mu = E(X) = p$. This can be shown using the definition of the expected value:

$$\begin{aligned} E(X) &= \sum_{x=0}^{1} k \cdot P(X = x) \\ &= 0 \cdot P(X = 0) + 1 \cdot P(X = 1) \\ &= p \end{aligned}$$

The second moment of the random variable can be found in a similar manner:

$$\begin{aligned} E(X^2) &= \sum_{X} x^2 P(X^2 = x) = \sum_{x=0}^{1} x^2 P(X^2 = x) \\ &= 0^2(1 - p) + 1^2(p) \\ &= p \end{aligned}$$

5.2.4 Variance

The variance of the Bernoulli(p) distribution is $\sigma^2 = p(1-p)$. This can be determined directly from the definition of the variance, $\sigma^2 = E(X^2) - \mu^2$. We've already calculated both the first and second moments of the random variable, so

$$\begin{aligned}\sigma^2 &= E(X^2) - [E(X)]^2 \\ &= p - p^2 \\ &= p(1-p)\end{aligned}$$

5.3 Binomial Distribution

The binomial distribution, or $X \sim \text{binomial}(n, p)$, is a discrete probability distribution of the number of successes with probability p in a sequence of n *independent* experiments, each asking a yes-no question. A binomial distribution with $n = 1$ events is also a Bernoulli distribution. Hence, a binomial distribution can be thought of as performing n independent Bernoulli(p) trials.

5.3.1 Probability Mass Function

The PMF of the binomial distribution is given by:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x \in \{0, 1, 2, \dots n\}$$

In order to be able to apply the binomial distribution, for a finite population of size n we must *replace* whatever is selected after each trial. If sampling is carried out without replacement, trials are not independent from one another and so the resulting distribution is *hypergeometric*. For an infinite population, replacement does not matter.

In practice, we could use a binomial distribution to model something like flipping a coin n times or rolling a dice n times. As an example, consider a biased coin which has a probability $p = 0.6$ of landing on heads and we decide to flip the coin 5 times. We can model this experiment with the random variable $X \sim \text{binomial}(5, 0.6)$. The probability of seeing exactly 2 heads in 5 flips is

$$\begin{aligned} P(X = 2) &= \binom{5}{2} (0.6)^2 (1 - 0.6)^3 \\ &\approx 0.2304 \end{aligned}$$

or about 23%.

5.3.2 Expected Value

The first moment of the random variable $X \sim \text{binomial}(n, p)$ is $\mu = E(X) = np$. This can be shown using the definition of the expected value:

$$\begin{aligned}
E(X) &= \sum_{x=0}^{n} x \cdot P(X = x) \\
&= \sum_{x=0}^{n} x \cdot \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} x \cdot \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{x \cdot n!}{x!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{n! \cdot p^x (1-p)^{n-x}}{(x-1)!(n-x)!} \\
&= np \sum_{x=1}^{n} \frac{(n-1)! \cdot p^{x-1} (1-p)^{(n-1)-(x-1)}}{(x-1)!(n-x)!} \\
&= np \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} (1-p)^{(n-1)-(x-1)} \\
&= np(p + 1 - p)^{n-1} \\
&= np
\end{aligned}$$

In the second to last step, we utilized the binomial theorem in order to simplify the expression. That is,

$$(a+b)^n = \sum_{i=1}^{n} \binom{n}{i} a^n b^{n-i}$$

This result follows from the linearity of the expected value along with the fact that X is the sum of n independent Bernoulli trials, each with expected value p. So, if $X_1, X_2, \ldots X_n$ are identical and independent Bernoulli variables with parameter p, then $X = X_1 + X_2 + \cdots + X_n$ and

$$E(X) = E(X_1) + E(X_2) + \cdots + E(X_n) = p + p + \cdots + p = np$$

The second moment of the random variable is $E(X^2) = np(np - p + 1)$. This can be found using the definition of the expected value:

$$\begin{aligned}
E(X^2) &= \sum_X x^2 P(X^2 = x) \\
&= \sum_{x=0}^{n} x^2 \binom{n}{x} p^x q^{n-x} \\
&= \sum_{x=0}^{n} \frac{n! x^2}{k!(n-x)!} p^x q^{n-x} \\
&= \sum_{x=0}^{n} \frac{n! x}{(x-1)!(n-x)!} p^x q^{n-x} \\
&= np \sum_{x=1}^{n} \frac{(n-1)! x}{(x-1)!(n-x)!} p^{x-1} q^{(n-1)-(x-1)} \\
&= np \sum_{x=1}^{n} x \binom{n-1}{x-1} p^{x-1} q^{(n-1)-(x-1)}
\end{aligned}$$

Here, let's make the substitution $j = x - 1$ and $m = n - 1$. Our expression then becomes:

$$\begin{aligned}
&= np \sum_{j=0}^{m} (j+1) \binom{m}{j} p^j q^{m-j} \\
&= np \left(\sum_{j=0}^{m} j \binom{m}{j} p^j q^{m-j} + \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} \right) \\
&= np \left(\sum_{j=0}^{m} m \binom{m-1}{j-1} p^j q^{m-j} + \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} \right)
\end{aligned}$$

The last transformation utilized the identity of the binomial coefficient:

$$j \binom{m}{j} = m \binom{m-1}{j-1}$$

Continuing,

$$\begin{aligned}
&= np \left(mp \sum_{j=1}^{m} \binom{m-1}{j-1} p^{j-1} q^{(m-1)-(j-1)} + \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} \right) \\
&= np[mp(p+q)^{m-1} + (p+q)^m] \\
&= np[(n-1)p(p+1-p)^{m-1} + (p+1-p)^m] \\
&= np[(n-1)p + 1] \\
&= n^2 p^2 - np^2 + np
\end{aligned}$$

Similar to the computation of the expected value, we utilized the binomial theorem to simplify the sums to binomials.

5.3.3 Variance

The variance of the binomial(n, p) distribution is $\sigma^2 = np(1-p)$. This can be found using the definition of the variance:

$$\begin{aligned}\sigma^2 &= E(X^2) - \mu^2 \\ &= n^2p^2 - np^2 + np - n^2p^2 \\ &= np - np^2 \\ &= np(1-p)\end{aligned}$$

We could have also arrived at this answer using the fact that a binomial(n, p) distribution is simply a sum of n independent Bernoulli(p) trials: The variance of a Bernoulli(p) trial is $\sigma^2 = p(1-p)$, and since each trial in a binomial distribution is independent of one another, we can simply sum the individual variances of each trial due to the linearity of the expected value. Let $Y \sim$ binomial(n, p) and $X \sim$ Bernoulli(p). Then,

$$\begin{aligned}\text{Var}(Y) &= \text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3) + \cdots + \text{Var}(X_n) \\ &= p(1-p) + p(1-p) + p(1-p) + \cdots + p(1-p) \\ &= np(1-p)\end{aligned}$$

which agrees with our formulation above.

5.4 Negative Binomial Distribution

The negative binomial distribution or Pascal distribution, denoted $X \sim \text{NB}(r, p)$, is a discrete probability distribution that models the number of successes in a sequence of independent and identically distributed Bernoulli trials before a specified number of successes occurs. This distribution answers the question, when does the rth success occur?

5.4.1 Probability Mass Function

The PMF of the negative binomial distribution is

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \quad \text{for } x \in \{r, r+1, r+2, \dots\}$$

where r is the number of successes, x is total number of trials, and p is the probability of success. To rationalize where this distribution comes from, consider the event that this distribution is modeling:

$$\begin{aligned} P(X = x) &= P[(r-1 \text{ successes in } x-1 \text{ trials}) \cap (\text{success in trial } x)] \\ &= P[(r-1 \text{ successes in } x-1 \text{ trials})] P[(\text{success in trial } x)] \\ &= \binom{x-1}{r-1} p^{r-1} (1-p)^{x-1-(r-1)} \cdot p \\ &= \binom{x-1}{r-1} p^r (1-p)^{x-r} \end{aligned}$$

Intuitively, the PMF for this distribution should make sense. For the first $k-1$ trials, we don't care which are successes and which are failures, all that matters is that the final trial is a success. After all, the negative binomial distribution answers the question of when the rth success occurs.

The negative binomial distribution can be thought of as the sum of independent geometric distributions: Consider the event that we are looking for 2 successes in a negative binomial distribution and let X_2 be the event of 2 successes. Then,

$$\begin{aligned} P(X_2 = x) &= \binom{x-1}{2-1} p^2 (1-p)^{x-2} \\ &= \binom{x-1}{1} p^2 (1-p)^{x-2} \\ &= (x-1) p^2 (1-p)^{x-2} \text{ for } x = 2, 3, 4, \dots \end{aligned}$$

Now, consider the sum of two independent geometric(p) random variables, Y_1 and Y_2. Then, using the law of total probability we have

$$P(Y_1 + Y_2 = x) = \sum_{j=1}^{\infty} P(Y_1 + Y_2 = x, Y_1 = j)$$

Here, we are "conditioning" on the event Y_1, that is, we are summing over all of the possible ways $Y_1 + Y_2 = x$ given that Y_1 occurred at the jth trial. Here, note that $x \geq 2$ because both $Y_1, Y_2 \geq 1, 2, 3, \dots$. Moreover, since we are assuming that Y_1 occurred before the xth trial, Y_1 is strictly less than x and we can stop our summation at $x-1$. Now, rearranging the interior of this probability,

$$P(Y_1 + Y_2 = x) = \sum_{j=1}^{x-1} P(Y_2 = x - j \cap Y_1 = j)$$

Since these two random variables Y_1 and Y_2 are independent by definition,

$$P(Y_1 + Y_2 = x) = \sum_{j=1}^{x-1} P(Y_2 = x - j)P(Y_1 = j)$$

Since Y_1 and Y_2 are defined to be geometric(p) distributions, we can substitute their probability mass functions into the summation:

$$\begin{aligned} P(Y_1 + Y_2 = x) &= \sum_{j=1}^{x-1} p(1-p)^{x-j-1} p(1-p)^{j-1} \\ &= \sum_{j=1}^{x-1} p^2(1-p)^{x-2} \\ &= (x-1)p^2(1-p)^{x-2} \\ &= P(X_2 = x) \end{aligned}$$

which is exactly our PMF for a negative binomial distribution! In the final step with the summation, there are no terms in the summand which depend on j and so we can simplify the summation as shown.

An example of the negative binomial distribution occurring in real life is when my sister Kate goes out to sell girl scout cookies. If she walks around our neighborhood of 30 houses where the probability of any individual house buying a single box of girl scout cookies is $p = 0.8$ and isn't supposed to stop until she's sold 10 boxes, what is the probability that she is done selling cookies at the 12th house she visits? (Obviously this is a contrived example; nobody buys just 1 box of girl scout cookies at a time). Here, we can model Kate's journey with the random variable $K \sim \text{NB}(10, 0.8)$. Thus, our calculation becomes:

$$\begin{aligned} P(K = 12) &= \binom{12-1}{10-1}(0.8)^{10}(1-0.8)^{12-10} \\ &= \binom{11}{9} 0.8^{10} 0.2^2 \\ &\approx 0.236 \end{aligned}$$

5.4.2 Expected Value

The expected value for a random variable $X \sim \text{NB}(r, p)$ distribution is $\mu = E(X) = \frac{r}{p}$. Using the definition of the expected value, this is difficult to show. Instead, we will take advantage of the fact that a negative binomial distribution is the sum of independent geometric(p) distributions. Let $X \sim \text{NB}(r, p)$ and $Y_1, Y_2, \ldots Y_n \sim$ geometric(p). Then,

$$\begin{aligned} E(X) &= E(Y_1 + Y_2 + \cdots + Y_n) \\ &= E(Y_1) + E(Y_2) + \cdots + E(Y_n) \\ &= \frac{1}{p} + \frac{1}{p} + \cdots + \frac{1}{p} \\ &= \frac{r}{p} \end{aligned}$$

5.4.3 Variance

The variance of the NB(r, p) distribution is $\sigma^2 = r(1-p)/p^2$.

5.5 Multinomial Distribution

The multinomial distribution is simply a generalization of the binomial distribution. For example, if we were to model the outcomes of a dice roll, a binomial distribution might model the probability distribution of rolling a 6 whereas a multinomial distribution could model the probability distribution of rolling each side. This generalization can also be seen in the *multinomial theorem*, which states that

$$(x_1 + x_2 + \cdots + x_r)^n = \sum_{a_1+a_2+\cdots+a_r=n} \binom{n}{a_1, a_2, \ldots, a_r} \prod_{i=1}^{r} x_i^{a_i}$$

where the multinomial coefficient is given by:

$$\binom{n}{a_1, a_2, \ldots, a_r} = \frac{n!}{a_1! a_2! \cdots a_r!}$$

5.5.1 Probability Mass Function

The PMF of $X \sim \text{multinomial}(n, r)$ is given by

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_r = x_r) = \binom{n}{x_1, x_2, \ldots, x_n} p_1^{x_1} p_2^{x_2} \cdots p_r^{x_r}$$

where n is the number of independent and identical trials where each trial can result in any one of r possible outcomes, each outcome having a probability p_i with $i = 1, 2, \ldots, r$. Furthermore, all $x_i \geq 0$, $\sum_{i=1}^{r} x_i = n$, and $\sum_{i=1}^{r} p_i = 1$. Two important facts about the multinomial are:

- The random variables X_i are dependent
- For each i, X_i has a binomial(n, p_i) distribution

The intuition behind the fact that each $X_i \sim \text{binomial}(n, p_i)$ can be shown by computing the marginal of X_i after summing out all other possible values of x_j, $j \neq i$. This would require some work. Another way of rationalizing this thought can be seen by fixing an $i \in \{1, 2, \ldots, r\}$. Then, letting X_i be the number of trials that result in the successful outcome i, $n - X_i$ becomes the number of trials that result in a failure of not i. The distribution of X_i becomes

$$P(X_i = \omega) = P(X_i = \omega, n - X_i = n - \omega) = \binom{n}{\omega} p_i^{\omega} (1 - p)^{n-\omega}$$

and X_i has a binomial distribution with parameters n and p_i.

Here's an example of the multinomial distribution in action: Suppose our friend Sean is running in a three-way election in the small town of Muscle Shoals, Alabama against two other candidates, we'll call them Digby and Frud. Sean received 50% of the votes, Digby 20% of the votes, and Frud 30% of the votes. If six voters are selected at random what is the probability that exactly one of them supports Digby, two of them support Frud, and three of them support Sean? Technically speaking this is a *multivariate hypergeometric distribution* since we are sampling without replacement. However, with a population of 14,103 (as of the 2019 census), we can assume the probabilities are unchanging from sample to sample. So, apply the PMF of the multinomial distribution:

$$P(\text{Sean} = 3, \text{Digby} = 1, \text{Frud} = 2) = \binom{6}{3, 1, 2} (0.5)^3 (0.2)^1 (0.3)^2 = 0.135$$

5.5.2 Expected Value

The expected value for any of the X_i's within a multinomial(n, r) distribution is $E(X_i) = np_i$. Intuitively this makes sense because we just noted that each of the X_i's within the multinomial follow a binomial(n, p_i) distribution.

5.5.3 Variance

The variance for any of the X_i's within the multinomial(n, r) distribution is $\text{Var}(X_i) = np_i(1 - p_i)$. Similar to the expected value, this is because each of the X_i's within the multinomial follow a binomial(n, p_i) distribution.

5.6 Geometric Distribution

The geometric distribution, or $X \sim \text{geometric}(p)$, is a discrete probability distribution that gives the probability that the first occurrence of a success requires k independent trials, each with a probability of success p.

A geometric distribution is appropriate for modeling if the following assumptions are true:

- The phenomenon being modeled is a sequence of independent trials
- There are only two possible outcomes for each trial, often designated as success or failure
- The probability of success p is the same for every trial

5.6.1 Probability Mass Function

The PMF can be described in one of two ways:

$$P(X = x) = p(1-p)^{x-1} \quad \text{for } x \in \{1, 2, 3 \dots\}$$
$$P(Y = y) = P(X = x + 1) = p(1-p)^{x} \quad \text{for } x \in \{0, 1, 2, \dots\}$$

where the random variable X (eq. 1) is used for modeling the number of trials up to and including the first success, while the random variable Y (eq. 2) models the number of failures *until* the first success. For the most part, this paper will use the random variable X instead of Y for the geometric distribution.

As a modeling tool, the geometric distribution arises in situations like flipping a coin until we see heads. As an example, let's pretend that I wanna have kids at 20 years old. What's the probability that I have 1 boy before having any girls? What about the probability of having 4 boys before having a girl? Let's also pretend that the probability of having a girl is $p = 0.6$. We can model this situation using the random variable $X \sim \text{geometric}(0.6)$. So, the probability that I have only 1 boy before having any girls is

$$\begin{aligned} P(X = 2) &= (0.6)(1 - 0.6)^1 \\ &= 0.24 \end{aligned}$$

Similarly, the probability that I have 4 boys before having a girl is given by

$$\begin{aligned} P(X = 5) &= (0.6)(1 - 0.6)^4 \\ &= 0.01536 \end{aligned}$$

A useful mathematical fact regarding sums is the sum for both finite and infinite geometric series. First, for a finite geometric series, the sum is given by:

$$\sum_{k=0}^{n} a_0 r^k = a_0 + a_0 r + a_0 r^2 + \cdots + a r^n = \frac{a_0(1 - r^{n+1})}{1 - r}$$

where a_0 is the first term in the series and $r \neq 1$ is the common ratio. For an infinite series,

$$\sum_{k=0}^{\infty} a_0 r^k = a_0 + a_0 r + a_0 r^2 + \cdots = \frac{a_0}{1 - r}$$

where $|r| < 1$ in order for the series to converge and for this expression to hold. These expressions will be used to evaluate the expected value of a geometric distribution later in the section.

5.6.2 Cumulative Distribution Function

The CDF of a geometric(p) distribution is given by

$$F(X) = 1 - (1-p)^x$$
$$F(Y) = 1 - (1-p)^{x+1}$$

5.6.3 Expected Value

The expected value for a geometric(p) distribution is $\mu = E(X) = \frac{1}{p}$. This can be shown many different ways:

Method 1: Definition of the Expected Value

$$\begin{aligned}
E(X) &= \sum_{x=1}^{\infty} x \cdot P(X = x) = \sum_{x=1}^{\infty} \sum_{j=1}^{x} P(X = x) = \sum_{j=1}^{\infty} \sum_{x=1}^{\infty} P(X = x) = \sum_{j=1}^{\infty} P(X \geq j) \\
&= \sum_{j=1}^{\infty} P(X = j) + P(X = j+1) + P(X = j+2) + \dots \\
&= \sum_{j=1}^{\infty} p(1-p)^{j-1} + p(1-p)^{j} + p(1-p)^{j+1} + \dots \\
&= \sum_{j=1}^{\infty} \frac{p(1-p)^{j-1}}{1-(1-p)} \\
&= \sum_{j=1}^{\infty} (1-p)^{j-1} \\
&= \frac{1}{1-(1-p)} \\
&= \frac{1}{p}
\end{aligned}$$

Method 2: Sum of Tail Probabilities

$$\begin{aligned}
E(X) &= \sum_{x=1}^{\infty} xP(X = x) = \sum_{x=1}^{\infty} xp(1-p)^{x-1} \\
&= p + 2p(1-p) + 3p(1-p)^2 + 4p(1-p)^3 + \dots \qquad (1)
\end{aligned}$$

Here, multiply both sides of the equation by $(1-p)$:

$$(1-p)E(X) = p(1-p) + 2p(1-p)^2 + 3p(1-p)^3 + 4p(1-p)^4 + \dots \qquad (2)$$

Now, subtract equation (2) from (1):

$$\begin{aligned}
E(X) - (1-p)E(X) = pE(X) &= p + p(1-p) + p(1-p)^2 + p(1-p)^3 + \dots \\
E(X) &= 1 + (1-p) + (1-p)^2 + (1-p)^3 + \dots
\end{aligned}$$

Now, we can calculate the expected value using the formula for the sum of an infinite series. Here, our common ratio is $1-p$ and the first term in the sequence is 1:

$$\begin{aligned}
E(X) &= 1 + (1-p) + (1-p)^2 + (1-p)^3 + \dots \\
&= \frac{1}{1-(1-p)} \\
&= \frac{1}{p}
\end{aligned}$$

which agrees with the answer from our other method.

Method 3: Calculus Argument

Lastly, my favorite method of finding the expected value for a geometric(p) distribution.

$$\begin{aligned}
E(X) &= \sum_{x=1}^{\infty} xP(X=x) = \sum_{x=1}^{\infty} xp(1-p)^{x-1} \\
&= p\sum_{x=1}^{\infty} x(1-p)^{x-1} \\
&= p\sum_{x=1}^{\infty} -\frac{d}{dp}(1-p)^{x} \\
&= -p\frac{d}{dp}\left(\sum_{x=1}^{\infty}(1-p)^{x}\right) \\
&= -p\frac{d}{dp}\left(\frac{1-p}{1-(1-p)}\right) = -p\frac{d}{dp}\left(\frac{1-p}{p}\right) \\
&= -p\frac{p(-1)-(1-p)}{p^2} \\
&= -p\frac{-1}{p^2} \\
&= \frac{1}{p}
\end{aligned}$$

which agrees with both of the other formulations for the expected value of a geometric(p) distribution. This method is sound because both differentiation and summation are linear operators.

5.6.4 Variance

The variance of a geometric(p) distribution is $\sigma^2 = (1-p)/p^2$. This can be found using the definition of the variance. Since we already know the mean of a geometric distribution, all we need to find is the second moment of the random variable:

$$\begin{aligned}
E(X^2) &= \sum_{X} x^2 P(X^2 = x) = \sum_{x=1}^{\infty} x^2 p(1-p)^{x-1} \\
&= p + 4p(1-p) + 9p(1-p)^2 + 16p(1-p)^3 + \ldots
\end{aligned}$$

Similar to method 2 of calculating the expected value, here we'll multiply both sides of this equation by $1-p$:

$$(1-p)E(X^2) = p(1-p) + 4p(1-p)^2 + 9p(1-p)^3 + 16p(1-p)^4 + \ldots$$

Now, subtract the new expression from the original:

$$\begin{aligned}
E(X^2) - (1-p)E(X^2) &= p + p(1-p) + 3p(1-p)^2 + 5p(1-p)^3 + 7p(1-p)^4 + \ldots \\
\iff pE(X^2) &= p + p(1-p) + 3p(1-p)^2 + 5p(1-p)^3 + 7p(1-p)^4 + \ldots
\end{aligned}$$

Here notice two things: 1) The coefficients of the terms containing $(1-p)$ are the odd natural numbers and 2) we can divide both sides of the equation by p to simplify:

$$E(X^2) = 1 + (1-p) + 3(1-p)^2 + 5(1-p)^3 + 7(1-p)^4 + \ldots$$

Now, let's rinse and repeat: Multiply both sides of this equation by $(1-p)$ and then subtract the new equation from the old one:

$$\begin{aligned}
(1-p)E(X^2) &= (1-p) + (1-p)^2 + 3(1-p)^3 + 5(1-p)^4 + 7(1-p)^5 + \dots \\
E(X^2) - (1-p)E(X^2) &= 1 + 2(1-p) + 2(1-p)^2 + 2(1-p)^3 + 2(1-p)^4 + \dots \\
pE(X^2) &= 1 + 2\sum_{i=1}^{\infty}(1-p)^i \\
&= 1 + 2\left(\frac{1-p}{1-(1-p)}\right) \\
&= 1 + \frac{2-2p}{p} = \frac{2-p}{p} \\
\iff E(X^2) &= \frac{2-p}{p^2}
\end{aligned}$$

Now, we can calculate the variance:

$$\begin{aligned}
\sigma^2 &= E(X^2) - \mu^2 \\
&= \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 \\
&= \frac{1-p}{p^2}
\end{aligned}$$

5.7 Hypergeometric Distribution

The hypergeometric distribution, or $X \sim \text{hypergeometric}(N, M, n)$, is a discrete probability distribution that describes the probability of x successes in n draws *without* replacement from a finite population of size N that contains exactly M favorable outcomes. In contrast, the binomial distribution describes drawing from a finite population *with* replacement.

The hypergeometric distribution is characterized by the following conditions:

- The result of each trial can be classified into one of two mutually exclusive categories, e.g., pass or fail, successful or unsuccessful.
- The probability of success changes on each trial, as each trial decreases the population

5.7.1 Probability Mass Function

The PMF of the hypergeometric distribution is given by:

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \quad \text{for } x = 0, 1, 2, \ldots, M$$

where N is the population size, M is the number of favorable outcomes, n is the number of trials where $0 \leq n \leq N$, and x is the number of observed successes where $0 \leq x \leq M$.

Hypergeometric distributions are very common in real life. Consider the following example which models the selecting of voters from a local district. Suppose we are interested in the small town of Big Sky, MT, which has 101 female and 95 male registered voters. We select a random sample of 10 voters. What is the probability that exactly 7 of them will be girls? We can model this situation using the random variable $X \sim \text{hypergeometric}(196, 101, 10)$. So, we can evaluate the probability that 7 females will be selected by:

$$\begin{aligned} P(X = 7) &= \frac{\binom{101}{7}\binom{95}{3}}{\binom{196}{10}} \\ &\approx 0.1304 \end{aligned}$$

5.7.2 Binomial Approximation to the Hypergeometric Distribution

The binomial approximation to the hypergeometric distributions posits that for large population sizes, the hypergeometric distribution can be estimated with a binomial distribution. So, let's show this. First consider the random variable $X \sim \text{Hypergeometric}(N, M, n)$ and suppose that n and x are fixed whereas N and M increase to infinity in such a way that $M/N \to p$ and therefore

$\frac{N-M}{M} \to 1-p$. Then, we can express the hypergeometric distribution in the following way:

$$\begin{aligned}
\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} &= \frac{M!}{x!(K-x)!} \cdot \frac{(N-M)!}{(n-x)!(N-M-(n-x))!} \cdot \frac{n!(N-n)!}{N!} \\
&= \binom{n}{x} \frac{M!}{(M-x)!} \cdot \frac{(N-M)!}{(N-M-(n-x))!} \cdot \frac{(N-n)!}{N!} \\
&\overset{*}{=} \binom{n}{x} \frac{M!(N-x)!}{(M-x)!N!} \cdot \frac{(N-M)!(N-n)!}{(N-x)!(N-M-(n-x))!} \\
&\overset{**}{=} \binom{n}{x} \frac{M!/(M-x)!}{N!/(N-x)!} \cdot \frac{(N-M)!/(N-M-(n-x))!}{(N-n+(n-x))!/(N-n)!} \\
&= \binom{n}{x} \prod_{i=1}^{x} \frac{M-x+i}{N-x+i} \prod_{j=1}^{n-x} \frac{N-M-(n-x)+j}{N-n+j}
\end{aligned}$$

*Here, we multiply by a form of 1. **Here, we rewrite $(N-x)! = (N+n-n-x)!$
Because we have that $M, N \to \infty$ and $M/N \to p$ for a fixed n and x, this expression affords the binomial PMF in the limit, since

$$\lim_{M,N\to\infty} \frac{M-x+i}{N-x+i} = \lim_{M,N\to\infty} \frac{M}{N} = p$$

and

$$\lim_{M,N\to\infty} \frac{N-M-(n-x)+j}{N-n+j} = \lim_{M,N\to\infty} \frac{N-M}{N} = 1-p$$

5.7.3 Expected Value

The first moment of the hypergeometric distribution is $\mu = E(X) = \frac{nM}{N}$. This can be shown using the definition of the expected value:

$$\begin{aligned}
E(X) &= \sum_{X} xP(X=x) \\
&= \sum_{x=0}^{n} \frac{x\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \\
&= \sum_{x=0}^{n} \binom{N-M}{n-x} \frac{xM!}{x!(M-x)!} \cdot \frac{n!(N-n)!}{N!}
\end{aligned}$$

Here, we'll start our sum at $x = 1$ instead of $x = 0$ because when $x = 0$, the expression evaluates to 0 and contributes nothing to the sum:

$$\begin{aligned}
\sum_{x=0}^{n} \binom{N-M}{n-x} \frac{xM!}{x!(M-x)!} \cdot \frac{n!(N-n)!}{N!} &= \sum_{x=1}^{n} \binom{N-M}{n-x} \frac{M(M-1)!}{(x-1)!(M-x)!} \cdot \frac{n(n-1)!(N-n)!}{N(N-1)!} \\
&= \sum_{x=1}^{n} \binom{N-M}{n-x} \frac{nM}{N} \cdot \frac{(M-1)!}{(x-1)!(M-x)!} \cdot \frac{(n-1)!(N-n)!}{(N-1)!} \\
&= \frac{nM}{N} \sum_{x=1}^{n} \frac{\binom{N-M}{n-x}\binom{M-1}{x-1}}{\binom{N-1}{n-1}} = \frac{nM}{N\binom{N-1}{n-1}} \sum_{x=1}^{n} \binom{N-M}{n-x}\binom{M-1}{x-1}
\end{aligned}$$

Here, we'll make the variable substitution $j = x - 1$ and therefore $x = j + 1$. By doing this, the sum will now begin at $j = 1 - 1 = 0$ and end at $n - 1$. Additionally, we'll utilize the combinatorial

identity

$$\sum_{y=0}^{m} \binom{a}{y}\binom{b}{m-y} = \binom{a+b}{m}$$

to simplify the summand:

$$\begin{aligned}\frac{nM}{N\binom{N-1}{n-1}} \sum_{x=1}^{n} \binom{N-M}{n-x}\binom{M-1}{x-1} &= \frac{nM}{N\binom{N-1}{n-1}} \sum_{j=0}^{n-1} \binom{N-M}{n-1-j}\binom{M-1}{j} \\ &= \frac{nM}{N\binom{N-1}{n-1}} \binom{N-1}{n-1} \\ &= \frac{nM}{N}\end{aligned}$$

which affords the expected value of the hypergeometric distribution.

5.7.4 Variance

The variance of the hypergeometric distribution is given by

$$\sigma^2 = \frac{nM}{N}\left(1 - \frac{M}{N}\right)\left(\frac{N-n}{N-1}\right)$$

Alternatively, if we define $p = M/N$ be the proportion of favorable outcomes in the entire population, we have

$$\sigma^2 = \left(\frac{N-n}{N-1}\right) np(1-p)$$

5.8 Poisson Distribution

The Poisson distribution, or $X \sim \text{Poisson}(\lambda)$, is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time if these events occur with a known constant mean rate and are independent of one another. The Poisson distribution can also be used for the number of events in other specified intervals, such as distance, area, and volume.

The Poisson distribution is an appropriate model if the following assumptions are true:

- x is the number of times an event occurs in an interval and x can take the values $0, 1, 2, \ldots$
- The occurrence of one event does not effect the probability that a second event will occur. That is, events occur independently
- The average rate at which events occur is independent of any occurrences. For simplicity, this is usually assumed to be constant, but may in practice vary with time
- Two events cannot occur at exactly the same instant; instead, at each very small sub-interval exactly one event either occurs or does not occur

5.8.1 Probability Mass Function

The PMF of the Poisson distribution is given by:

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{for } x \in \{0, 1, 2, \ldots\}$$

Here, X counts the number of "events" that occur in a given amount of *exposure* (time, space, etc.) and λ is the *rate* of exposure, with $\lambda > 0$, representing the expected number of "events" in the exposure.

Poisson distributions arise often in everyday life, modeling things from the number of patients entering the emergency room between 11pm and midnight to the number of photons hitting a detector in a particular time interval. As an example, consider the number of category 5 hurricanes that made landfalls in the United States in the past century. According to Dr. Torcaso, 38 of these have caused landfalls. Thus, for our Poisson distribution, $\lambda = 38/100 = 0.38$. Now, what is the probability a cat-5 hurricane makes a landfall in the next two years? We can model this situation with the random variable $X \sim \text{Poisson}(0.76)$. Why does $\lambda = 0.76$? Remember that λ is the average number of events happening within a certain period of exposure. Since we assumed that rate of cat-5 hurricanes causing landfalls was a constant 0.38 and we are considering the exposure over the course of 2 years, $\lambda = 0.38 \cdot 2 = 0.76$. Since we are looking for the probability that a cat-5 hurricane makes a landfall in the next two years, we are looking for the probability

$$\begin{aligned} P(X \geq 1) &= 1 - P(X = 0) \\ &= 1 - \frac{e^{-0.76}(0.76)^0}{0!} \\ &\approx 0.5323 \end{aligned}$$

Consider a second example involving the number of goals scored in an individual World Cup match. Some reports cite that the average number of goals scored in a game is approximately 2.5 and that the Poisson model is appropriate. Because the average event rate is 2.5 goals per match,

$\lambda = 2.5$. From here we can compute the probability of scoring X goals in a match:

$$\begin{aligned} P(X = 1 \text{ goal}) &= \frac{e^{-2.5}(2.5)^1}{1!} \\ &\approx 0.2052 \\ P(X = 2 \text{ goals}) &= \frac{e^{-2.5}(2.5)^2}{2!} \\ &\approx 0.2565 \\ P(X = 8 \text{ goals}) &= \frac{e^{-2.5}(2.5)^8}{8!} \\ &\approx 0.0031 \end{aligned}$$

For context, the probability that the 2014 World Cup final would contain 8 scored goals was $> 1\%$. I'm not saying it was rigged, but...

5.8.2 Poisson Limit Theorem

The Poisson distribution can be thought of as a limit of binomials. Suppose that the fixed amount of exposure we care about is on the unit interval $[0, 1]$. Now, break the interval into n parts, where n is relatively large. Let $Y \sim \text{binomial}(n, \frac{\lambda}{n})$ (alternatively, we could express the distribution as $Y \sim \text{Bernoulli}(\frac{\lambda}{n})$) where n is the number of trials we'll run and $\frac{\lambda}{n}$ is the probability each small segment of the unit interval has of being selected. So, the random variable Y counts the number of successes that occur within the interval. Now, using the probability mass function of the binomial distribution,

$$\begin{aligned} P(Y = y) &= \binom{n}{y} \left(\frac{\lambda}{n}\right)^y \left(1 - \frac{\lambda}{n}\right)^{n-y} \\ &= \frac{n!}{y!(n-y)!} \cdot \frac{\lambda^y}{n^y} \cdot \left(1 - \frac{\lambda}{n}\right)^{n-y} \\ &= \frac{\lambda^y (1 - \frac{\lambda}{n})^{n-y}}{y!} \cdot \frac{n(n-1)(n-2)\dots(n-1-(y-1))}{n^y} \\ &= \frac{\lambda^y (1 - \frac{\lambda}{n})^{n-y}}{y!} \cdot \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\dots\left(1 - \frac{y-1}{n}\right) \end{aligned}$$

Since y is a fixed value and we let n be arbitrarily large, we can take the limit as $n \to \infty$ for the expression above:

$$\begin{aligned} &= \lim_{n\to\infty} \frac{\lambda^y (1 - \frac{\lambda}{n})^{n-y}}{y!} \cdot \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\dots\left(1 - \frac{y-1}{n}\right) \\ &= \frac{\lambda^y}{y!} \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n \\ &= \frac{e^{-\lambda}\lambda^y}{y!} \end{aligned}$$

And we recover our equation for the PMF of the Poisson distribution. In the final step we utilized the known limit of e,

$$e = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n$$

and alternatively, in the form of a sum,

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

Both of these expressions are important for calculating the expected value and variance of the Poisson distribution.

Another interesting consequence of the Poisson distribution is that the sum of two independent Poisson random variables is another Poisson random variable. Consider the two random variables $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poissson}(\lambda_2)$ and let $S = X_1 + X_2$. What is the PMF of S? To do this, we can use the law of total probability to express the event S in terms of X_1 and X_2:

$$P(S = x) = P(X_1 + X_2 = x) = \sum_{y=0}^{x} P(X_1 + X_2 = x, X_2 = y)$$

With this substitution, we are finding the probability of the event $P(X_1 + X_2 = x)$ for some particular $X_2 = y$, and then summing those particular probabilities over all possible values for y on $0 \leq y \leq x$. Substituting $X_2 = y$ into the first expression in the sum affords $X_1 = x - y$. Thus, our summand simplifies to:

$$P(S = x) = \sum_{y=0}^{x} P(X_1 = x - y, X_2 = y)$$

Because it is given to us that X_1 and X_2 are independent random variables, we can further simplify this expression to

$$P(S = x) = \sum_{y=0}^{x} P(X_1 = x - y)P(X_2 = y)$$

allowing us to substitute the PMFs for X_1 and X_2 and evaluate the sum:

$$\begin{aligned}
P(S = x) &= \sum_{y=0}^{x} \left(\frac{e^{-\lambda_1}\lambda_1^{x-y}}{(x-y)!}\right)\left(\frac{e^{-\lambda_2}\lambda_2^{y}}{y!}\right) \\
&= e^{-(\lambda_1+\lambda_2)} \sum_{y=0}^{x} \frac{\lambda_1^{x-y}\lambda_2^{y}}{y!(x-y)!} \\
&= e^{-(\lambda_1+\lambda_2)} \sum_{y=0}^{x} \frac{\lambda_1^{x-y}\lambda_2^{y}}{y!(x-y)!} \cdot \frac{x!}{x!} \\
&= e^{-(\lambda_1+\lambda_2)} \sum_{y=0}^{x} \binom{x}{y} \frac{\lambda_1^{x-y}\lambda_2^{y}}{x!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^{x}}{x!} = \frac{e^{-\lambda_S}\lambda_S^{x}}{x!}
\end{aligned}$$

And we recover the PMF for S. The parameter for S, λ_S, is simply the sum of the parameters of X_1 and X_2.

One aspect of the Poisson distribution to note is that for large values of n, that is, for a large number of independent trials, Poisson(np) approximates binomial(n, p) very well.

5.8.3 Expected Value

The expected value for the Poisson distribution is $\mu = E(X) = \lambda$, the parameter of the function. This can be shown using the definition of the expected value:

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x \cdot P(X = x) \\ &= \sum_{x=1}^{\infty} \frac{x e^{-\lambda} \lambda^x}{x!} \\ &= \sum_{x=1}^{\infty} \frac{e^{-\lambda} \lambda^x}{(x-1)!} \\ &= e^{-\lambda} \lambda^x \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\ &= e^{-\lambda} \lambda e^{\lambda} \\ &= \lambda \end{aligned}$$

5.8.4 Variance

The variance of the Poisson(λ) distribution is $\sigma^2 = \lambda$. To show why this is true, we'll first utilize the linearity of the expected value to rewrite the second moment of the random variable in a more computationally friendly way:

$$E(X^2) = E(X^2 - X + X) = E[X(X-1) + X] = E[X(X-1)] + E(X)$$

The term $E[X(X-1)]$ is known as the second factorial moment of the random variable X. Now, let's compute the second factorial moment of the random variable:

$$E[X(X-1)] = \sum_{x=0}^{\infty} x(x-1) P[X(X-1) = x] = \sum_{x=0}^{\infty} x(x-1) \left(\frac{e^{-\lambda} \lambda^x}{x!} \right)$$

Notice that for both $x = 0$ and $x = 1$, the expression evaluates to 0 and contributes nothing to the sum. So, we can start our summation at $x = 2$ without changing the value of the sum.

$$\begin{aligned} E[X(X-1)] &= \sum_{x=2}^{\infty} x(x-1) \left(\frac{e^{-\lambda} \lambda^x}{x!} \right) \\ &= \sum_{x=2}^{\infty} \frac{e^{-\lambda} \lambda^x}{(x-2)!} \\ &= e^{-\lambda} \lambda^2 \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} \\ &= e^{-\lambda} \lambda^2 e^{\lambda} \\ &= \lambda^2 \end{aligned}$$

Interestingly enough, the nth factorial moment of the Poisson distribution is λ^n. Now, let's compute the second moment of the random variable:

$$E(X^2) = E[X(X-1)] + E(X) = \lambda^2 + \lambda$$

And finally, the variance:

$$\begin{aligned}\sigma^2 &= E(X^2) - \mu^2 = \lambda^2 + \lambda - \lambda^2 \\ &= \lambda\end{aligned}$$

An interesting aspect of the Poisson(λ) distribution is that Var(X)=E(X).

5.9 Logarithmic Distribution

5.9.1 Probability Mass Function

5.9.2 Expected Value

5.9.3 Variance

6 Continuous Random Variables

6.1 Continuous Uniform Distribution

The *continuous uniform distribution* (or *rectangular distribution*) is a family of symmetric probability distributions that describes an experiment where there is an outcome which lies between certain bounds where all intervals of the same length on the distribution are equally likely to occur.

6.1.1 Probability Density Function

The PDF of the uniform distribution, denoted $X \sim \mathcal{U}(a, b)$ for a continuous random variable X, is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

Graphically, the PDF is portrayed as a rectangle with a base on the interval $b - a$ and a height of $\frac{1}{b-a}$.

Consider the following example that utilizes the continuous uniform distribution: After going to Menards I have two loaves of bread A and B in my fridge of lengths a and b, respectively, where $a < b$. Now, I decide to cut loaf B at random uniformly along its length. What is the probability that the 3 resulting loaves of bread form a triangle? As mentioned in the question, this probability will follow a uniform distribution. In particular, for a random variable X, we have that $X \sim \mathcal{U}(0, b)$. Thus, the PDF of X is given by

$$f_X(x) = \frac{1}{b} I_{[0,b]}$$

where $I_{[0,b]}$ is an indicator function. When we cut loaf B, we acquire two new pieces of bread of lengths l and $b - l$. Recall the following about triangles: In order to create a triangle, the sum of any two sides must be greater than the length of the third side. In terms of inequalities for our situation with loaves of lengths a, l, and $b - l$,

$$\begin{aligned} a + l &> b - l \\ a + b - l &> l \\ b - l + l &> a \end{aligned}$$

Using these inequalities, we can deduce the following about the side length l:

$$\frac{b-a}{2} < l < \frac{a+b}{2}$$

Thus, we are looking for the probability mass of X on the interval which l is contained. So, integrate

the PDF of X using these bounds:

$$\begin{aligned}
P\left(\frac{b-a}{2} < X < \frac{a+b}{2}\right) &= \int_{\frac{b-a}{2}}^{\frac{a+b}{2}} \frac{1}{b} I_{[0,b]}\, dx \\
&= \frac{1}{b}\int_{\frac{b-a}{2}}^{\frac{a+b}{2}} dx \\
&= \frac{1}{b}\left[x\right]_{\frac{b-a}{2}}^{\frac{a+b}{2}} \\
&= \frac{1}{b}\left[\frac{a+b}{2} - \frac{b-a}{2}\right] \\
&= \frac{a}{b}
\end{aligned}$$

Thus, the probability of being able to form a triangle after I haphazardly slice these loaves of bread is $\frac{a}{b}$.

6.1.2 Cumulative Distribution Function

The cumulative distribution function of $X \sim \mathcal{U}(a, b)$ is given by

$$F_X(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{1}{b-a}(x-a) & \text{for } a \leq x \leq b \\ 1 & \text{for } b < x \end{cases}$$

This result can be arrived at easily using the definition of the CDF.

6.1.3 Expected Value

The first moment of $X \sim \mathcal{U}(a, b)$ is $E(X) = \frac{1}{2}(b + a)$. This can be determined using the definition of the expected value. Note that the support of this distribution is defined to be on $[a, b]$, allowing us to simplify the constants of integration:

$$\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x f_X(x)\, dx \\
&= \int_a^b \frac{k}{b-a}\, dx = \frac{1}{b-a}\left[\frac{1}{2}x^2\right]_a^b \\
&= \frac{1}{b-a}\left[\frac{1}{2}(b^2 - a^2)\right] = \frac{1}{3(b-a)}(b+a)(b-a) \\
&= \frac{1}{2}(b+a)
\end{aligned}$$

The second moment of the random variable can be found in a similar manner:

$$\begin{aligned}
E(X^2) &= \int_{-\infty}^{\infty} x^2 f_X(x)\, dx \\
&= \int_a^b \frac{x^2}{b-a}\, dx = \frac{1}{b-a}\left[\frac{1}{3}x^3\right]_a^b \\
&= \frac{1}{b-a}\left[\frac{1}{3}(b^3 - a^3)\right] = \frac{1}{3(b-a)}(b-a)(b^2 + ab + a^2) \\
&= \frac{b^2 + ab + a^2}{3} = \frac{b^3 - a^3}{3(b-a)}
\end{aligned}$$

Either expression in the last line can be cited as the second moment.

In general, the nth moment of the uniform distribution is given as:

$$E(X^n) = \frac{b^{n+1} - a^{n+1}}{(n-1)(b-a)} = \frac{1}{n+1} \sum_{k=0}^{n} a^x b^{n-x}$$

6.1.4 Variance

The variance of $X \sim \mathcal{U}(a, b)$ is $\sigma^2 = \text{Var}(X) = \frac{1}{12}(b-a)^2$. This can be found using the definition of the variance:

$$\begin{aligned}
\text{Var}(X) &= E(X^2) - [E(X)]^2 \\
&= \frac{b^3 - a^3}{3(b-a)} - \left(\frac{1}{2}(b+a)\right)^2 \\
&= \frac{b^3 - a^3}{3(b-a)} - \frac{1}{4}(b+a)^2 = \frac{4(b^3 - a^3) - 3(b+a)^2(b-a)}{12(b-a)} \\
&= \frac{4(b^3 - a^3) - 3(b^3 + ab^2 - a^2 b - a^3)}{12(b-a)} \\
&= \frac{b^3 + 3ab^2 - 3a^2 b - a^3}{12(b-a)} = \frac{(b-a)^3}{12(b-a)} \\
&= \frac{1}{12}(b-a)^2
\end{aligned}$$

6.1.5 Moment Generating Function

The moment generating function of $X \sim \mathcal{U}(a, b)$ is given by

$$M(x) = E(e^{\theta x}) = \frac{e^{\theta b} - e^{\theta a}}{\theta(b-a)}$$

This can be found directly using the definition of the expected value and the LOTUS:

$$\begin{aligned}
E(e^{\theta x}) &= \int_{-\infty}^{\infty} e^{\theta k} f_X(x)\, dx = \int_a^b e^{\theta x} \frac{1}{b-a}\, dx \\
&= \frac{1}{b-a} \int_a^b e^{\theta x}\, dx = \frac{1}{b-a} \left[\frac{1}{\theta} e^{\theta x}\right]_a^b \\
&= \frac{1}{\theta(b-a)} \left[e^{\theta b} - e^{\theta a}\right] \\
&= \frac{e^{\theta b} - e^{\theta a}}{\theta(b-a)}
\end{aligned}$$

6.2 Exponential Distribution

The exponential distribution can be imagined as the probability distribution of the time between events in a Poisson point process, i.e., a process in which events occur continuously and independently at a constant and average rate. It is a particular case of the gamma distribution.

6.2.1 Probability Density Function

The PDF of $X \sim \text{Exp}(\lambda)$ is given by:

$$f_X(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0$$

λ is often referred to as the *rate parameter* of the distribution.

6.2.2 Cumulative Distribution Function

The CDF of $X \sim \text{Exp}(\lambda)$ is given by:

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

6.2.3 Expected Value

The first moment of $X \sim \text{Exp}(\lambda)$ is $E(X) = \frac{1}{\lambda}$. This can be shown using the definition of the expected value. Recall that the support of the exponential distribution is $x \geq 0$:

$$E(X) = \int_{-\infty}^{\infty} x\lambda e^{-\lambda x}\, dx$$

We'll integrate this with integration by parts, using

$$\begin{aligned} u &= x, \quad du = dx \\ dv &= \lambda e^{-\lambda x}, \quad v = -e^{-\lambda x} \end{aligned}$$

and substituting these values affords

$$\begin{aligned} \int_{-\infty}^{\infty} x\lambda e^{-\lambda x}\, dx &= \Big[-xe^{-\lambda x} \Big]_0^{\infty} + \int_0^{\infty} e^{-\lambda u}\, du \\ &= 0 + \left[\frac{-1}{\lambda} e^{-\lambda u} \right]_0^{\infty} \\ &= \frac{1}{\lambda} \end{aligned}$$

The second moment of $X \sim \text{Exp}(\lambda)$ is $E(X^2) = \frac{2}{\lambda^2}$ and can be computed in a similar manner as above:

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \lambda e^{-\lambda x}\, dx = \int_0^{\infty} x^2 \lambda e^{-\lambda x}\, dx$$

Here, I'm going to use a result from the section on the gamma function to save myself from performing integration by parts like I'm some pleb:

$$\begin{aligned} \int_0^{\infty} x^2 \lambda e^{-\lambda x}\, dx &= \lambda \left(\frac{1}{\lambda} \right)^3 \Gamma(3) \\ &= \frac{\lambda}{\lambda^3}(2!) = \frac{2}{\lambda^2} \end{aligned}$$

6.2.4 Variance

The variance of $X \sim \text{Exp}(\lambda)$ is $\text{Var}(X) = \frac{1}{\lambda^2}$. This can be shown using the definition of the variance:

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

6.2.5 Moment Generating Function

The moment generating function $X \sim \text{Exp}(\lambda)$ is

$$M(x) = \frac{\lambda}{\lambda - \theta} \quad \text{for } \theta < \lambda$$

This can be shown using the definition of the moment generating function. Recall that the support of the exponential distribution is $0 \to \infty$:

$$\begin{aligned} E(e^{\theta X}) &= \int_{-\infty}^{\infty} e^{\theta x}(\lambda e^{-\lambda x})\, dx = \int_0^{\infty} \lambda e^{x(\theta-\lambda)}\, dx \\ &= \lambda \left[\frac{1}{\theta - \lambda} e^{x(\theta-\lambda)}\right]_0^{\infty} \end{aligned}$$

Here, note that in order for this integral to converge, it must be the case that $\theta < \lambda$. Placing these restrictions, we have that

$$E(e^{\theta X}) = \frac{\lambda}{\lambda - \theta} \quad \text{for } \theta < \lambda$$

6.2.6 Memorylessness Property

The memorylessness property of the exponential distribution states that for a random variable $X \sim \text{Exp}(\lambda)$, for all $s, t > 0$, then

$$P(X > s + t \mid X > t) = P(X = s)$$

The proof of this statement can be shown using the definition of the conditional probability:

$$P(X > s + t \mid X > t) = \frac{P(X > s + t, X > t)}{P(X > t)}$$

Since $P(X > t)$ is a subset of $P(X > s + t)$, the intersection of their probabilities can be simplified to $P(X > s + t, X > t) = P(X > s + t)$:

$$\begin{aligned} \frac{P(X > s + t, X > t)}{P(X > t)} &= \frac{P(X > t + s)}{P(X > t)} \\ &= \frac{\lambda e^{-\lambda(s+t)}}{\lambda e^{-\lambda t}} \\ &= e^{-\lambda s} = P(X > s) \end{aligned}$$

The memorylessness property helps us to answer questions like, “Given that a lightbulb has lasted 2 years already, what is the probability that it will last another 3 years?" As long as this lightbulb’s lifetime adheres to an exponential distribution, we can ignore the conditioning “2 years" and just consider the probability that it will last 3 years.

6.3 Normal (Gaussian) Distribution

The Normal Distribution is, quite possibly, the most important probability distribution that humans have ever come across. No less of a man than Carl Gauss himself did us the courtesy of studying and characterizing this distribution. Normal distributions are used often in natural and social sciences to represent real-valued variables whose distributions are unknown. Another aspect of their character which makes them so important is in part due to the *Central Limit Theorem*. The central limit theorem states that in many situations, when independent random variables are added, their (adequately normalized) sum tends towards a normal distribution even if the original variables themselves are not normally distributed. This theorem is a key concept in probability and statistics because it implies that the normal distribution can be applicable in many situations involving other types of distributions.

6.3.1 Probability Density Function

The probability density function of a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ is given by

$$f_X(x) = \frac{e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}}{\sigma\sqrt{2\pi}} \quad \text{for } -\infty < x < \infty$$

where μ is the mean of the distribution and σ is the standard deviation. Are we sure that this is a PDF? Yes. To show why, we have to prove that the integral of the function over all reals evaluates to 1. So, consider the following integral:

$$I = \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}}{\sigma\sqrt{2\pi}} dx$$

For the sake of simplification, we'll let $u = (x - \mu)\sigma$ and therefore $du = \sigma dx$ to afford

$$I = \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}u^2}}{\sqrt{2\pi}} du$$

If you've taken a calculus class, you've probably already recognized that an integral of this form cannot afford a closed solution. While this is true for *definite* integrals, this is not necessarily true for *indefinite* integrals, such as the one we have here. So how can we go about solving? First, square both sides of the equation:

$$\begin{aligned} I^2 = \left(\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}u^2}}{\sqrt{2\pi}} du \right)^2 &= \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}u^2}}{\sqrt{2\pi}} du \cdot \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}v^2}}{\sqrt{2\pi}} dv \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}(u^2+v^2)}}{2\pi} du dv \end{aligned}$$

After squaring both sides we've created a situation where our result for I could be either positive *or* negative. However, since the function we are integrating is nonnegative on all of $\mathbb{R}$, we do not need to worry about I being negative because our function is necessarily positive. From here, we can make a variable substitution into *polar coordinates*, allowing $r^2 = u^2 + v^2$. With this substitution, we have new constants of integration. Since r is defined to be positive, $0 \leq r < \infty$. Also, since u and v can take on any value in the real numbers, θ must range from $0 \leq \theta \leq 2\pi$. Additionally, recall that the Jacobian for a transformation into polar coordinates is r. Thus, our new double integral becomes

$$I^2 = \int_0^{2\pi} \int_0^{\infty} \frac{e^{-\frac{1}{2}r^2}}{2\pi} r dr d\theta$$

which can be evaluated to afford a solution:

$$\begin{aligned} I^2 &= \int_0^{2\pi} \int_0^{\infty} \frac{re^{-\frac{1}{2}r^2}}{2\pi} dr d\theta \\ &= \frac{1}{2\pi} \int_0^{2\pi} d\theta \int_0^{\infty} re^{-\frac{1}{2}r^2} \\ &= \frac{1}{2\pi} \Big[\theta\Big]_0^{2\pi} \Big[-e^{-\frac{1}{2}r^2}\Big]_0^{\infty} \\ &= \frac{1}{2\pi}(2\pi)(1) \\ &= 1 \end{aligned}$$

With this, we've shown that $I^2 = 1$ and therefore $I = 1$ because we've already ruled out the integral evaluating to -1 due to the fact that the function is necessarily positive (this is mentioned earlier). Thus, we've shown that the normal distribution is in fact a PDF.

A useful property of the normal distribution is as follows: If $X \sim \mathcal{N}(\mu, \sigma^2)$ and a, b are any constants with $a \neq 0$, then

$$Y = aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$

Now, we go on to prove this result. Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$ and let a, b be constants with $a \neq 0$. Now, define a random variable Y such that $Y = aX + b$. The CDF of Y is given by

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(aX + b \leq y) \\ &= P(aX \leq y - b) = P\left(X \leq \frac{y-b}{a}\right) \end{aligned}$$

We'll first consider the case where $a > 0$ (the case illustrated above). Then, by the definition of the CDF, we have

$$F_Y(y) = \int_0^{\frac{y-b}{a}} \frac{e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2}}{\sigma\sqrt{2\pi}} dt$$

where the integrand is the PDF of Y. By the fundamental theorem of calculus, we can find the PDF of Y:

$$\begin{aligned} f_Y(y) = \frac{d}{dy} F_Y(y) &= \frac{d}{dy} \int_0^{\frac{y-b}{a}} \frac{e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2}}{\sigma\sqrt{2\pi}} dt \\ &= \frac{e^{-\frac{1}{2}(\frac{\frac{y-b}{a}-\mu}{\sigma})^2}}{\sigma\sqrt{2\pi}} \left(\frac{d}{dy}\left[\frac{y-b}{a}\right]\right) \end{aligned}$$

Note that in the last step, we multiply by the derivative of $(y - b)/a$ due to the chain rule. Simplifying this integral,

$$\frac{e^{-\frac{1}{2}(\frac{\frac{y-b}{a}-\mu}{\sigma})^2}}{\sigma\sqrt{2\pi}} \left(\frac{d}{dy}\left[\frac{y-b}{a}\right]\right) = \frac{e^{-\frac{1}{2}(\frac{y-(a\mu+b)}{a\sigma})^2}}{a\sigma\sqrt{2\pi}}$$

Notice that the above expression is exactly the form of the PDF for a random variable $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$! (not a factorial)

Extending from this result is what's known as the *Z-score transformation*: If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = aX + b$, with $a = \frac{1}{\sigma}$ and $b = -\frac{\mu}{\sigma}$, then we have the following relationships:

$$\begin{cases} a\mu + b = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0 \\ a^2\sigma^2 = \frac{\sigma^2}{\sigma^2} = 1 \\ aX + b = \frac{X}{\sigma} - \frac{\mu}{\sigma} = \frac{X-\mu}{\sigma} \end{cases}$$

Now, we define a random variable $Z = \frac{X-\mu}{\sigma}$, called the z-score transformation, to X to attain the *standard normal distribution*. Everything we just did involved *random variable transforms*, a topic of the utmost importance in probability and statistics and discussed more in a later section.

Standard Normal Distribution

We refer to the $\mathcal{N}(0, 1)$ distribution as the standard normal distribution. It has a PDF given by

$$\varphi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}} \quad \text{for } -\infty < x < \infty$$

and a CDF given by

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt$$

These functions get their own variables because, like the normal distribution itself, they're special. Typically, we reserve the letter Z to denote a random variable which has a standard normal distribution.

6.3.2 Cumulative Distribution Function

The CDF of the normal distribution has no closed form. For $X \sim \mathcal{N}(\mu, \sigma^2)$, the CDF is given by the integral

$$F_X(x) = \int_{-\infty}^{x} \frac{e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2}}{\sigma\sqrt{2\pi}}\, dt$$

where the integrand is simply the PDF of the normal distribution.

6.3.3 Expected Value

The expected value of $X \sim \mathcal{N}(\mu, \sigma^2)$ is $E(X) = \mu$.

6.3.4 Variance

The variance of $X \sim \mathcal{N}(\mu, \sigma^2)$ is $\text{Var}(X) = \sigma^2$.

6.4 Log-Normal Distribution

The log-normal distribution is a probability distribution of a random variable whose logarithm normally distributed. Thus, it is often times written as $Y = e^X$ where X is normally distributed. The log-normal distribution shows up in a lot of niche places. Topics ranging from the length of comments posted on an internet discussion forum, to certain physiological measures such as blood pressure in adults, to the molar mass distribution of linear polymers.

6.4.1 Probability Density Function

If a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, then the PDF of $Y = e^X$ is given by

$$f_Y(y) = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\ln y - \mu}{\sigma})^2} \quad \text{for } y > 0$$

A simple question we could ask knowing that $Y = e^X$ is log-normally distributed is, what is the distribution of cY for some real parameter $c > 0$? What about $c + Y$? We can answer both of these using the variable transform and ignoring the cumbersome PDF. First, cY: Since $Y = e^X$,

$$cY = ce^X = e^X e^{\ln c} = e^{\ln c + X}$$

and because $X \sim \mathcal{N}(\mu, \sigma^2)$, it follows that $X + \ln c \sim \mathcal{N}(\mu + \ln c, \sigma^2)$ and therefore Y is log-normally distributed with parameters $\ln c + \mu$ and σ^2. Now $c + Y$: Since Y is log-normally distributed, it must be the case that $P(Y \leq y) > 0$ for all $y > 0$ (this is a direct result of the restrictions on y in the PDF of the log-normal distribution). Therefore, if $P(Y \leq y) = 0$ it follows that Y *cannot* be log-normally distributed. Now, suppose $W = c + Y$. The CDF of W evaluated at c is given by

$$P(W \leq c) = P(c + Y \leq c) = P(Y \leq 0) = 0$$

and because $c > 0$, it is impossible for W to be log-normally distributed. Thus, $c + Y$ does not follow a log-normal distribution.

6.4.2 Expected Value

The expected value of $Y = e^X$ where $X \sim \mathcal{N}(\mu, \sigma^2)$ is $E(Y) = e^{\mu + \frac{\sigma^2}{2}}$.

6.4.3 Variance

The variance of $Y = e^X$ where $X \sim \mathcal{N}(\mu, \sigma^2)$ is $\text{Var}(X) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$.

6.5 Gamma Distribution

The Gamma distribution is really cool primarily because it involves the gamma function, $\Gamma(\alpha)$.

6.5.1 Gamma Function

The gamma function, $\Gamma(\alpha)$, is the most commonly used extension of the factorial function for complex numbers. The gamma function is defined for all complex numbers except for negative integers. By definition, the gamma function is the improper integral

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y}\, dy \text{ for } \mathrm{Re}(x) > 0$$

The gamma function is a *transcendental function*, meaning that is cannot be structurally simplified (to a polynomial, for example). Consider an example to illustrate a standard calculation using the gamma function: Evaluate $\Gamma(1)$.

$$\begin{aligned}\Gamma(1) &= \int_0^\infty y^{1-1} e^{-y}\, dy = \int_0^\infty e^{-y}\, dy \\ &= \Big[e^{-y}\Big]_0^\infty \\ &= 1\end{aligned}$$

Thus, $\Gamma(1) = 1$.

An incredible result of the gamma function is that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. We can show this is true with (relative) ease by using the definition of the gamma function:

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty y^{\frac{1}{2}-1} e^{-y}\, dy = \int_0^\infty y^{-\frac{1}{2}} e^{-y}\, dy$$

Now, let's make the variable substitution $y = \frac{u^2}{2}$ and therefore $dy = udu$. This affords

$$\begin{aligned}\int_0^\infty y^{-\frac{1}{2}} e^{-y}\, dy &= \int_0^\infty \left(\frac{u^2}{2}\right)^{-\frac{1}{2}} e^{-\frac{u^2}{2}}\, udu \\ &= \sqrt{2}\int_0^\infty e^{-\frac{u^2}{2}}\, du\end{aligned}$$

This integral looks *almost* like that of the PDF for the standard normal distribution. So, let's multiply by a form of 1 to get it to look like that. To do this, we'll multiply by $\frac{\sqrt{2\pi}}{\sqrt{2\pi}}$:

$$\left(\frac{\sqrt{2\pi}}{\sqrt{2\pi}}\right)\sqrt{2}\int_0^\infty e^{-\frac{u^2}{2}}\, du = 2\sqrt{\pi}\int_0^\infty \frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}}\, du$$

Now, we're left with the standard normal distribution evaluated from $0 \to \infty$. Since the normal distribution is symmetric about its mean, we know that this integral evaluates to 0. Thus,

$$2\sqrt{\pi}\int_0^\infty \frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}} = 2\sqrt{\pi}\left(\frac{1}{2}\right) = \sqrt{\pi}$$

which returns the claimed result.

Reduction Property

The reduction property of the gamma function states that

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$$

We can illustrate this proof with ease. Consider some $\alpha > 1$:

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y}\, dy$$

To begin to evaluate this integral we'll use integration by parts, with,

$$u = y^{\alpha-1}, \quad du = (\alpha - 1)y^{\alpha-2}$$
$$dv = e^{-y}, \quad v = -e^{-y}$$

which affords

$$\begin{aligned}\int_0^\infty y^{\alpha-1} e^{-y}\, dy &= \Big[- y^{\alpha-1} e^{-y} \Big]_0^\infty + \int_0^\infty (\alpha - 1) y^{\alpha-2} e^{-y}\, dy \\ &= 0 + (\alpha - 1) \int_0^\infty y^{\alpha-2} e^{-y}\, dy \\ &= (\alpha - 1)\Gamma(\alpha - 2)\end{aligned}$$

We could continue integrating this integral by parts to illustrate the general result, or I could just state it. For any integer n,

$$\Gamma(n) = (n - 1)!$$

We can illustrate the usefulness of the reduction property with the following example: Evaluate $\Gamma(\frac{7}{2})$. We could go about using the definition of the gamma function and start taking integrals, but that's no fun. Instead, we'll use the reduction property of the gamma function:

$$\begin{aligned}\Gamma\left(\frac{7}{2}\right) &= \left(\frac{7}{2} - 1\right)\Gamma\left(\frac{7}{2} - 1\right) = \left(\frac{5}{2}\right)\Gamma\left(\frac{5}{2}\right) = \left(\frac{5}{2}\right)\left(\frac{5}{2} - 1\right)\Gamma\left(\frac{5}{2} - 1\right) \\ &= \left(\frac{5}{2} \cdot \frac{3}{2}\right)\Gamma\left(\frac{3}{2}\right) \\ &= \frac{5}{2} \cdot \frac{3}{2} \cdot \frac{1}{2}\Gamma\left(\frac{1}{2}\right) \\ &= \frac{5}{2} \cdot \frac{3}{2} \cdot \frac{1}{2}(\sqrt{\pi}) = \frac{15\sqrt{\pi}}{8}\end{aligned}$$

Here, I'll remark that if we know (or can approximate) the value of $\Gamma(x)$ for some $0 < x \leq 1$, then we can compute $\Gamma(\alpha)$ for *any* $\alpha > 0$ via the reduction property.

6.5.2 Probability Density Function

The PDF of $X \sim \text{Gamma}(\alpha, \beta)$ is given by

$$f_X(x) = \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} \quad \text{for } x > 0$$

where $\alpha > 0$ is known as the *shape parameter* and $\beta > 0$ is known as the *scale parameter*. To show where this PDF comes from, in the words of Bill Nye, consider the following: Let $\alpha > 0$ and $\beta > 0$ and suppose we have the integral:

$$\int_0^\infty y^{\alpha-1} e^{-y/\beta}\, dy$$

We'll use a u-sub to begin to evaluate this, with

$$u = \frac{y}{\beta}, \quad \beta du = dx$$

which affords

$$\begin{aligned}\int_0^\infty y^{\alpha-1} e^{-y/\beta}\, dy &= \int_0^\infty (\beta u)^{\alpha-1} e^{-u}\, \beta du \\ &= \beta^\alpha \int_0^\infty u^{\alpha-1} e^{-u}\, du \\ &= \beta^\alpha \Gamma(\alpha)\end{aligned}$$

Thus, we have that

$$\bigstar \int_0^\infty y^{\alpha-1} e^{-y/\beta}\, dy = \beta^\alpha \Gamma(\alpha) \bigstar$$

Before talking about how this relates to the PDF, look! We now have a way to easily calculate those dumb integration by parts questions that we had to do over and over again in Calc in High School! Step aside with your contrived integration by parts examples, Mr. Bridger, I have no need for you anymore (I'm kidding of course, you're the best :)). Let's look at the following example to show the power of this formula. Evaluate

$$\int_0^\infty x^3 e^{-x/2}\, dx$$

Normally, we'd bust out the u's and v's and start integrating by parts. However, now that we have the equation above we can skip the hassle and plug in α and β directly:

$$\int_0^\infty x^3 e^{-x/2}\, dx = \int_0^\infty x^{4-1} e^{-x/2}\, dx = (2^4)\Gamma(4) = 16(3!) = 96$$

Now, how does this relate to the PDF? Since we've already shown that,

$$\int_0^\infty y^{\alpha-1} e^{-y/\beta}\, dy = \beta^\alpha \Gamma(\alpha)$$

dividing by $\beta^\alpha \Gamma(\alpha)$ on either side will *normalize* the integral, affording a PDF. That is:

$$\frac{\int_0^\infty y^{\alpha-1} e^{-y/\beta}\, dy}{\beta^\alpha \Gamma(\alpha)} = 1$$

This expression affords the PDF of the gamma distribution.

6.5.3 Expected Value

The expected value of $X \sim \text{Gamma}(\alpha, \beta)$ is $E(X) = \alpha\beta$. The first moment of the gamma distribution can be computed using the definition of the expected value for a continuous random

variable. Suppose $X \sim \text{Gamma}(\alpha, \beta)$. Also, recall that the support for the gamma distribution is the positive reals. Then,

$$\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x \cdot \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)}\, dx = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int_{0}^{\infty} x^{\alpha} e^{-x/\beta}\, dx \\
&= \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int_{0}^{\infty} x^{(\alpha+1)-1} e^{-x/\beta}\, dx \\
&= \frac{1}{\beta^{\alpha}\Gamma(\alpha)} (\beta^{\alpha+1}\Gamma(\alpha+1)) \\
&= \frac{\beta^{\alpha+1}(\alpha)\Gamma(\alpha)}{\beta^{\alpha}\Gamma(\alpha)} \\
&= \alpha\beta
\end{aligned}$$

The second moment of X can be computed in a similar manner:

$$\begin{aligned}
E(X^2) &= \int_{-\infty}^{\infty} x^2 \cdot \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)}\, dx = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int_{0}^{\infty} x^{\alpha+1} e^{-x/\beta}\, dx \\
&= \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int_{0}^{\infty} x^{(\alpha+2)-1} e^{-x/\beta}\, dx \\
&= \frac{1}{\beta^{\alpha}\Gamma(\alpha)} (\beta^{\alpha+2}\Gamma(\alpha+2)) \\
&= \frac{\beta^{\alpha+2}(\alpha+1)(\alpha)\Gamma(\alpha)}{\beta^{\alpha}\Gamma(\alpha)} \\
&= \alpha(\alpha+1)\beta^2
\end{aligned}$$

6.5.4 Variance

The variance for $X \sim \text{Gamma}(\alpha, \beta)$ is $\text{Var}(X) = \alpha\beta^2$. This can be shown using the definition of the variance:

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \alpha(\alpha+1)\beta^2 - (\alpha\beta)^2 = \alpha\beta^2$$

6.5.5 Moment Generating Function

The moment generating function of $X \sim \text{Gamma}(\alpha, \beta)$ is given by

$$E(e^{\theta X}) = \frac{1}{(1-\theta\beta)^{\alpha}} \quad \text{for } \theta < \frac{1}{\beta}$$

We can show this using the Law of the Unconscious Statistician:

$$\begin{aligned}
E(e^{\theta X}) &= \int_{-\infty}^{\infty} e^{\theta x} \left(\frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \right) dx \\
&= \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int_{0}^{\infty} x^{\alpha-1} e^{-x(\frac{1}{\beta}-\theta)}\, dx \\
&= \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int_{0}^{\infty} x^{\alpha-1} e^{-x/[\beta/(1-\theta\beta)]}\, dx \\
&= \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \left[\left(\frac{\beta}{1-\theta\beta} \right)^{\alpha} \Gamma(\alpha) \right] \\
&= \left(\frac{\beta}{\beta(1-\theta\beta)} \right)^{\alpha} \\
&= \frac{1}{(1-\theta\beta)^{\alpha}}
\end{aligned}$$

6.5.6 Gamma Distribution Connection to Other Distributions

The Exponential Distribution as a Class of The Gamma Distribution

Consider $X \sim \text{Gamma}(1, \frac{1}{\lambda})$. The PDF of this distribution is given by

$$f_X(x) = \frac{x^{1-1}e^{-\lambda x}}{(1/\beta)^1\Gamma(1)} = \lambda e^{-\lambda x}$$

which is exactly the PDF of the exponential distribution. Thus, we see that $\text{Gamma}(1, \frac{1}{\lambda}) = \text{Exp}(\lambda)$.

Relation to The Poisson Distribution

The Chi-square Distribution as a Class of The Gamma Distribution

Consider $X \sim \text{Gamma}(\frac{\nu}{2}, 2)$. The PDF of this distribution is given by

$$f_X(x) = \frac{x^{\frac{\nu}{2}-1}e^{-x/2}}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})}$$

which is exactly the PDF of the chi-square distribution. Thus, we see that $\text{Gamma}(\frac{\nu}{2}, 2) = \chi^2(\nu)$ with $\nu > 0$.

6.6 Beta Distribution

The Beta distribution is a family of distributions defined on the interval $[0, 1]$ characterized by two positive real parameters, α and β.

6.6.1 Probability Density Function

The PDF of $X \sim \text{Beta}(\alpha, \beta)$ is given by

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad \text{for } 0 \le x \le 1$$

This is also sometimes written as:

$$f_X(x) = \frac{1}{\text{B}(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where B is the beta function, given by

$$\text{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

6.6.2 Expected Value

The expected value of $X \sim \text{Beta}(\alpha, \beta)$ is $E(X) = \frac{\alpha}{\alpha+\beta}$. This can be shown using the definition of the expectation:

$$\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{(\alpha+1)-1}(1-x)^{\beta-1} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{(\alpha+1)-1}(1-x)^{\beta-1} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(\frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)} \right) \\
&= \frac{\alpha}{\alpha+\beta}
\end{aligned}$$

The second moment of this distribution can be found in a similar manner:

$$\begin{aligned}
E(X^2) &= \int_{-\infty}^{\infty} x^2 f_X(x)\, dx = \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{(\alpha+2)-1}(1-x)^{\beta-1} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{(\alpha+2)-1}(1-x)^{\beta-1} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(\frac{\Gamma(\alpha+2)\Gamma(\beta)}{\Gamma(\alpha+2+\beta)} \right) \\
&= \frac{\alpha(\alpha+1)}{(\alpha+\beta+1)(\alpha+\beta)}
\end{aligned}$$

which is about as simplified as we could care to make this expression.

6.6.3 Variance

The variance of $X \sim \text{Beta}(\alpha, \beta)$ is $\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. We can illustrate this result quite easily as follows:

$$\begin{aligned}
\text{Var}(X) &= E(X^2) - [E(X)]^2 \\
&= \frac{\alpha(\alpha+1)}{(\alpha+\beta+1)(\alpha+\beta)} - \left(\frac{\alpha}{\alpha+\beta}\right)^2 \\
&= \frac{\alpha(\alpha+1)(\alpha+\beta)}{(\alpha+\beta+1)(\alpha+\beta)^2} - \frac{\alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)} \\
&= \frac{\alpha^3 + \alpha^2\beta + \alpha^2 + \alpha\beta - \alpha^3 - \alpha^2\beta - \alpha^2}{(\alpha+\beta+1)(\alpha+\beta)^2} \\
&= \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}
\end{aligned}$$

6.7 Other Distributions I Think Are Cool

6.7.1 Cauchy Distribution

The PDF of $X \sim \text{Cauchy}(\alpha, \gamma)$ is given by

$$f_X(x) = \frac{1}{\pi\gamma}\left[\frac{\gamma^2}{(x-\alpha)^2 - \gamma^2}\right] \quad \text{for } -\infty < x < \infty$$

whereα is the *location parameter* and γ is the *scale parameter*.

6.7.2 Weibull Distribution

The PDF of $X \sim \text{Weibull}(\lambda, k)$ is given by

$$f_X(x) = \frac{ke^{-(x/\lambda)^k}}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \quad \text{for } x \geq 0$$

where $k > 0$ is the *shape parameter* and $\lambda > 0$ is the *scale parameter*.

6.7.3 Rayleigh Distribution

The PDF of $X \sim \text{Rayleigh}(\sigma)$ is given by

$$f_X(x) = \frac{x}{\sigma^2}e^{-x^2/2\sigma^2} \quad \text{for } x \geq 0$$

where σ is the *scale parameter*.

6.7.4 Pareto Distribution

The PDF of $X \sim \text{Pareto}(\alpha, \lambda)$ is given by

$$f_X(x) = \frac{\alpha\lambda^\alpha}{x^{\alpha+1}} \quad \text{for } x \geq \lambda$$

where α is the *shape parameter* and λ is the *scale parameter*.

6.7.5 Chi-Squared Distribution

The PDF of $X \sim \chi^2_n$ is given by

$$f_X(x) = \frac{x^{(n/2)-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)} \quad \text{for } x > 0$$

where n are the *degrees of freedom* of the distribution.

6.7.6 F-Distribution

The PDF of $X \sim \text{F}_{n,m}$ is given by:

$$f_X(x) = \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}\left(\frac{n}{m}\right)^{\frac{n}{2}}\left(1+\frac{nx}{m}\right)^{-\frac{1}{2}(n+m)} x^{\frac{n}{2}-1} \quad \text{for } x > 0$$

More commonly, the F-distribution is expressed as a ratio of chi-squared distributions: If $U \sim \mathrm{F}_{n,m}$ and $X \sim \chi^2_n$ and $Y \sim \chi^2_m$, then

$$U = \frac{X/n}{Y/m}$$

6.7.7 Student's t-Distribution

The PDF of $X \sim$ Student's t_m is given by:

$$f_X(x) = \frac{\Gamma(\frac{m+1}{2})}{\sqrt{m\pi}\Gamma(\frac{m}{2})}\left(1+\frac{x^2}{m}\right)^{-\frac{1}{2}(m+1)} \quad \text{for } x > 0$$

The t-distribution is also often expressed in terms of a variable transform. If $X \sim$ Student's t_m and $Z \sim \mathcal{N}(0,1)$, $Y \sim \chi^2_n$ and Z and Y are independent, then

$$X = \frac{Z}{\sqrt{\frac{Y}{n}}}$$

6.7.8 Double Exponential (Laplace) Distribution

The PDF of $X \sim \mathcal{L}(\lambda)$ is given by:

$$f_X(x) = \frac{\lambda}{2}e^{-\lambda(x-c)} \quad \text{for } x, c > \mathbb{R} \text{ and } \lambda > 0$$

where c is the *location parameter* and λ is the *scale parameter*.

7 Jointly Distributed Random Variables

Given an experiment with a sample space Ω we may be able to define many random variables. Consider the following two random variables:

$$X : \Omega \to \mathbb{R} \quad , \quad Y : \Omega \to \mathbb{R}$$

When considering whether or not two variables qualify for being jointly distributed they must both exist in the same sample space, in this case Ω. Since X and Y meet this requirement, we'll call them *jointly distributed*. X and Y can be discrete random variables, continuous random variables, or a combination of both.

food for thought 16.51

7.0.1 The Jointly Discrete Case

If two random variables X and Y are jointly distributed discrete random variables, they have a joint PMF of

$$P_{X,Y}(x, y) = P(X = x, Y = y) = P(\{X = x\} \cap \{Y = y\})$$

defined on the support of X and Y such that

$$\sum_X \sum_Y P(X = x, Y = y) = 1$$

Consider this very simple example to illustrate the idea of the jointly discrete case: Suppose I have a jar of 3 red, 4 white, and 5 blue marbles and I select two marbles without replacement and take note of the colors. Let X represent the number of red marbles I select and Y represent the number of blue I select. Note that we can represent these variables as sets, with $X = \{0, 1, 2\}$ and $Y = \{0, 1, 2\}$. Since we don't a PMF to express these jointly distributed variables, we'll elect to use a table to portray all of the possible probabilities of this experiment:

$P(X = i, Y = j)$	$Y = 0$	$Y = 1$	$Y = 2$	PMF of X
$X = 0$	6/66	20/66	10/66	36/66
$X = 1$	12/66	15/66	0	27/66
$X = 2$	3/66	0	0	3/66
PMF of Y	21/66	35/66	10/66	

But Sam, how did you compute any of these? Great question. First we'll talk about those probabilities which are zero. Recall that we are only drawing 2 balls. Therefore, any probability which involves 3 or more balls must be zero because it will never happen. Now for the nonzero probabilities. These were computed using basic counting principles. Say we draw 1 red marble and 1 blue marble. There are $\binom{3}{1}$ ways to draw the red marble and $\binom{5}{1}$ ways to choose the blue marble. In total, there are $\binom{12}{2}$ ways to draw marbles. So, the probability of drawing 1 red marble and 1 blue marble is

$$P(X = 1, Y = 1) = \frac{\binom{3}{1}\binom{5}{1}}{\binom{12}{2}} = \frac{3 \cdot 5}{66} = \frac{15}{66}$$

Similarly, we can compute the probability of drawing neither a red nor a blue marble in either selection:

$$P(X = 0, Y = 0) = \frac{\binom{4}{2}}{\binom{12}{2}} = \frac{6}{66}$$

An identical procedure can be carried out for the rest of the probabilities in this table. The bottom row and rightmost column, labeled "PMF of Y" and "PMF of X", respectively, are referred to as the *marginal distributions*. Notice that the marginal distributions of both X and Y sum to 1 (as they should, because the marginal distributions represent the distributions of X and Y by themselves). Using this table, we can compute some more interesting probabilities: What is the probability that $X = Y$?

$$P(X = Y) = \frac{6}{66} + \frac{15}{66} + 0 = \frac{21}{66}$$

What is the probability that $X^2 + Y^2 \leq 1$?

$$P(X^2 + Y^2 \leq 1) = \frac{6}{66} + \frac{12}{66} + \frac{20}{66} = \frac{38}{66}$$

(In this case, I omitted all of the cases of zero probability). What is the probability that $X \leq 1$?

$$P(X \leq 1) = \frac{27}{66} + \frac{36}{66} = \frac{63}{66}$$

Notice that we can find the univariate probability of X even though it is jointly distributed with Y by using the marginal distribution of X.

7.0.2 The Jointly Continuous Case

If X and Y are jointly distributed continuous random variables, they have a joint PDF $f_{X,Y}(x, y) : \mathbb{R}^2 \to [0, \infty]$ such that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dA = 1$$

We'll illustrate this idea with an example: Consider two jointly distributed continuous random variables X and Y with the following joint distribution:

$$f_{X,Y}(x, y) = \begin{cases} e^{-y} & \text{for } 0 < x < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

First, let's show that this function is a PDF. Clearly, $f_{X,Y} \geq 0$ on all of $\mathbb{R}^2$ because it is an exponential function. So, we're left to show that it integrates to 1. Notice that the support of the function is on $x \in (0, y)$ and $y \in (0, \infty)$:

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dA &= \int_0^{\infty} \int_0^{y} e^{-y}\, dx dy \\ &= \int_0^{\infty} y e^{-y}\, dy = \Gamma(2) = 1 \end{aligned}$$

My life has improved drastically since learning about the gamma function solely because I don't have to do integration by parts for these kinds of integrals anymore. With that, we've shown that $f_{X,Y}(x, y)$ is a PDF.

Now, let's use this PDF to compute the joint probability $P(X + Y \leq 1)$:

$$\begin{aligned} P(X+Y \leq 1) &= \iint_{X+Y \leq 1} f_{X,Y}(x,y)\, dA \\ &= \int_0^{\frac{1}{2}} \int_x^{1-x} e^{-y}\, dy dx \\ &= \int_0^{\frac{1}{2}} \Big[-e^{-y} \Big]_x^{1-x} dx \\ &= \int_0^{\frac{1}{2}} -e^{x-1} + e^{-x}\, dx \\ &= \Big[-e^{x-1} - e^{-x} \Big]_0^{\frac{1}{2}} \\ &= -2e^{-\frac{1}{2}} + e^{-1} + 1 \end{aligned}$$

Ta-da! It's like magic except without the illusion part.

7.0.3 The Mixed Case

If X and Y are jointly distributed where X is a discrete random variable and Y is a continuous random variable, they have a joint distribution $f_{X,Y}(x,y)$ such that

$$\sum_X \left[\int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy \right] = \int_{-\infty}^{\infty} \left[\sum_X f_{X,Y}(x,y) \right] dy = 1$$

Suppose we have two random variables N and X with the following joint distribution:

$$P_{N,X}(n,x) = n\left(\frac{1}{2}\right)^n e^{-nx} \quad \text{for } n = 1,2,3, x > 0$$

Here, it is possible that N may be a discrete random variable and X may be a continuous random variable. In particular, this distribution is a PMF in n for every fixed choice of x and is a PDF in x for every fixed choice of n. To apply this joint distribution, imagine we play a game where we flip a coin repeatedly and observe the trial n of the first head where $N \sim$ geometric($\frac{1}{2}$). Then, we win \$$X$ where $X \sim \text{Exp}(n)$. We'll come back to playing with this distribution soon when we talk about marginal distributions (like the very next thing). For now, however, we'll let the mixed joint distribution be.

7.1 Marginal Distributions

The marginal distributions of jointly distributed random variables are the probability distributions of the individual random variables without reference to the values of any other variables. Consider a joint distribution $f_{X,Y}(x,y)$. In the discrete case, the marginal distributions of X and Y are given by

$$f_X(x) = \sum_Y f_{X,Y}(x,y) \quad , \quad f_Y(y) = \sum_X f_{X,Y}(x,y)$$

respectively. In the continuous case, the result is immensely similar:

$$f_X(x) = \int_Y f_{X,Y}(x,y)\, dy \quad , \quad f_Y(y) = \int_X f_{X,Y}(x,y)\, dx$$

7.2 Conditional Distributions

Let X and Y be jointly continuous random variables with joint PDF $f_{X,Y}(x, y)$. The conditional distribution of X given $Y = y$ is given by,

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

which is the joint distribution over the marginal distribution of Y. Similarly, the conditional distribution of Y given $X = x$ is:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

Notice that from the above definitions,

$$f_{X,Y}(x, y) = f_{Y|X}(y|x)f_X(x) = f_{X|Y}(x|y)f_Y(y)$$

revealing that the joint PDFs can be constructed from the conditional distribution and the corresponding marginal distribution.

Notice again (I know, I'm asking a lot) that when X and Y are jointly continuous,

$$P(X \in A|Y = y) = \int_A f_{X|Y}(x|y)\, dx$$

even thought the *probability* of the event $Y = y$ is zero because Y is a continuous random variable. In contrast, if we try to compute $P(X \in A|Y \in B)$ with $P(Y \in B) > 0$, then:

$$P(X \in A|Y \in B) = \frac{\iint_{B,A} f_{X,Y}(x, y)\, dxdy}{\int_B f_Y(y)\, dy}$$

In general, however,

$$P(X \in A|Y \in B) \neq \iint_{B,A} f_{X|Y}(x|y)\, dxdy$$

i.e., don't use the conditional density distribution to compute the conditional probability.

Here's an example illustrating the conditional distribution: Suppose we have two random variables X and Y with the joint PDF $f(x, y) = e^{-y}$ for $0 < x < y < \infty$. Firstly, compute the marginal distribution of Y. We can do this by integrating out the contribution from X in the joint PDF:

$$f_Y(y) = \int_0^y e^{-y}\, dx = ye^{-y} \quad \text{for } y > 0$$

Secondly, compute the probability $P(X \geq 1|Y = 2)$. To do this, we'll use the definition of the conditional distribution. First, find the conditional distribution of $X|Y = y$:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{e^{-y}}{ye^{-y}} = \frac{1}{y} \quad \text{for } 0 < x < y$$

Now, we can use the conditional distribution to compute the conditional probability:

$$P(X \geq 1|Y = 2) = \int_X f_{X|Y}(x|y)\, dx = \int_1^y \frac{1}{y}\, dx = \int_1^2 \frac{1}{2}\, dx = \frac{1}{2}$$

Amazing! Lastly, compute the probability $P(X \geq 1|Y \geq 2)$. Instead of using the conditional distribution, let's do this using the integral of the joint distribution over the integral of the marginal distribution of Y:

$$\begin{aligned} P(X \geq 1|Y \geq 2) &= \frac{\iint_{B,A} f_{X,Y}(x,y)\,dxdy}{\int_B f_Y(y)\,dy} = \frac{\int_2^\infty \int_1^y e^{-y}\,dxdy}{\int_2^\infty ye^{-y}\,dy} \\ &= \frac{\int_2^\infty (y-1)e^{-y}\,dy}{[-ye^{-y}]_2^\infty + \int_2^\infty e^{-y}\,dy} = \frac{\int_2^\infty ye^{-y} - e^{-y}\,dy}{2e^{-2} + [-e^{-y}]_2^\infty} \\ &= \frac{\int_2^\infty ye^{-y}\,dy - \int_2^\infty e^{-y}\,dy}{2e^{-2} + e^{-2}} \\ &= \frac{3e^{-2} - e^{-2}}{3e^{-2}} = \frac{2}{3} \end{aligned}$$

Here's another example illustrating the conditional distribution: Suppose $X|\Theta \sim \text{binomial}(n, \theta)$ where $\Theta \sim \mathcal{U}(0,1)$. Derive the joint distribution of X and Θ, the marginal distribution of X, and the conditional distribution of $\Theta|X$. We'll do all this stuff in order. Recall from above that we can find the joint distribution using the product of the conditional distribution and the distribution of the conditioning variable. That is,

$$\begin{aligned} P_{X,\Theta}(x,\theta) &= P_{X|\Theta}(x|\theta) f_\Theta(\theta) \\ &= \left(\binom{n}{x}\theta^x(1-\theta)^{n-x}\right)\left(\frac{1}{1-0}\right) \\ &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \end{aligned}$$

where $0 < \theta < 1$ and $x = 0, 1, 2, \ldots, n$. The marginal distribution of X can be found by integrating out the contribution from Θ:

$$\begin{aligned} P(X = x) &= \int_0^1 \binom{n}{x}\theta^x(1-\theta)^{n-x}\,d\theta \\ &= \binom{n}{x}\int_0^1 \theta^x(1-\theta)^{n-x}\,d\theta \end{aligned}$$

Notice that this integral is an un-normalized Beta(α, β) distribution with $\alpha = x + 1$ and $\beta = n - x + 1$. So, it simplifies to:

$$\begin{aligned} \binom{n}{x}\int_0^1 \theta^x(1-\theta)^{n-x}\,d\theta &= \binom{n}{x}\left(\frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)}\right) \\ &= \frac{n!}{x!(n-x)!}\left(\frac{x!(n-x)!)}{(n+1)n!}\right) = \frac{1}{n+1} \quad \text{for } x = 0, 1, 2, \ldots, n \end{aligned}$$

Recalling the identity $\Gamma(n) = (n-1)!$ is important for doing some of that. Now we have the

distribution of X. Lastly, let's compute conditional distribution of $\Theta|X$:

$$\begin{aligned} f_{\Theta|X}(\theta|x) &= \frac{P_{X,\Theta}(x,\theta)}{P_X(x)} = \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x}}{\frac{1}{n+1}} \\ &= \left(\frac{n!}{x!(n-x)!}\right)(n+1)\theta^x(1-\theta)^{n-x} \\ &= \left(\frac{(n+1)!}{x!(n-x)!}\right)\theta^x(1-\theta)^{n-x} \\ &= \left(\frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)}\right)\theta^x(1-\theta)^{n-x} \end{aligned}$$

This is *another* Beta$(x+1, n-x+1)$ distribution! One thing to note that doesn't have anything to do with the problem but is just interesting is that in statistics, Θ would be known as the *prior distribution* and $\Theta|X$ the *posterior distribution*.

7.3 Expected Value

The joint expected value of two random variables is computed no differently than when the random variables are not jointly distributed. Just like in the univariate case, the Law of the Unconscious Statistician holds for jointly distributed random variables. In symbols,

$$\begin{aligned} E(g(X,Y)) &= \sum_Y \sum_X g(X,Y)P(X=x, Y=y) \quad \text{if } X, Y \text{ are jointly discrete} \\ &= \iint_{X,Y} g(X,Y)f_{X,Y}(x,y)\,dxdy \quad \text{if } X, Y \text{ are jointly continuous} \end{aligned}$$

By linearity of the expectation, if X and Y are independent random variables, then

$$E(g(X)h(Y)) = E(g(X))E(h(Y))$$

for two real-valued functions g and h where $g(X)$ and $h(Y)$ have finite means. This can be shown with relative ease:

$$\begin{aligned} E[g(X)h(Y)] &= \iint_{X,Y} g(X)h(Y)f_{X,Y}(x,y)\,dxdy \\ &= \iint_{X,Y} g(X)h(Y)f_X(x)f_Y(y)\,dxdy \\ &= \int_X g(X)f_X(x)\,dx \int_Y h(Y)f_Y(y)\,dy \\ &= E[g(X)]E[h(Y)] \end{aligned}$$

Inductively, it follows that for independent random variables $X_1, \ldots, X_n$ and a real-valued function g,

$$E\left(\prod_{i=1}^n g(X_i)\right) = \prod_{i=1}^n E[g(X_i)]$$

Here is an example calculation of the joint expectation: Suppose $X \sim \text{Exp}(1)$ and $Y \sim \mathcal{U}(0,1)$ are independent random variables. Compute $E\left(e^{-XY^2}\right)$. This expectation is given by

$$E\left(e^{-XY^2}\right) = \iint_{X,Y} e^{-xy^2} f_{X,Y}(x,y)\,dxdy$$

and so let's first find the joint PDF of X and Y. Since they are independent,

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) = (e^{-x})(1) = e^{-x} \quad \text{for } x > 0, y \in [0,1]$$

which allows us to simplify the integral for the expectation and solve:

$$\begin{aligned} E\left(e^{-XY^2}\right) &= \int_0^1 \int_0^\infty e^{-xy^2} e^{-x}\, dx dy \\ &= \int_0^1 \left[-\frac{1}{1+y^2} e^{-x(1+y^2)} \right]_0^\infty dy \\ &= \int_0^1 \frac{1}{1+y^2}\, dy \\ &= \Big[\arctan y \Big]_0^1 = \frac{\pi}{4} \end{aligned}$$

Additionally, we can begin to discuss the *conditional expectation* when dealing with jointly distributed variables. We define the conditional expectation of two random variables X and Y as follows:

$$\begin{aligned} E(X|Y=y) &= \sum_X P(X=x|Y=y) \quad \text{if } X \text{ is discrete} \\ &= \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx \quad \text{if } X \text{ is continuous} \end{aligned}$$

Notice that $E(X|Y) = g(Y)$ can be thought of as a function of the random variable $Y = y$. There are six important properties regarding the conditional expectation, the first being the definition mentioned above. Here are four more:

- For two jointly distributed variables X and Y and some function $h : \mathbb{R} \to \mathbb{R}$,

$$E[h(Y)X|Y] = h(Y)E(X|Y)$$

- For some jointly distributed random variables U, V, W and constants a, b,

$$E(aU + bV|W) = aE(U|W) + bE(V|W)$$

- If X and Y are independent random variables,

$$E(X|Y) = E(X)$$

- If c is a constant,

$$E(c|Y) = c$$

The last property I refrained from mentioning because it's very powerful. Here it is:

7.3.1 Law of Total Expectation

By definition, the Law of Total Expectation states that

$$E[E(X|Y)] = E(X)$$

A proof of this fact can be done for both the discrete, continuous, and mixed jointly distributed cases. Here are the joint discrete and joint continuous cases: Let X and Y be jointly distributed discrete random variables. Then,

$$\begin{aligned} E[E(X|Y)] = E[g(Y)] &= \sum_Y g(y)P(Y=y) \\ &= \sum_Y \sum_X xP(X=x|Y=y)P(Y=y) \end{aligned}$$

Here, we've used $g(Y) = E(X|Y)$ as a random variable in order to compute this expectation. In the second step, we simply substituted the expectation of $E(X|Y) = y$. We could have done this without using $g(Y)$, but Dr. Torcaso did it in lecture so I did it here. Continuing,

$$\begin{aligned} \sum_Y \sum_X xP(X=x|Y=y)P(Y=y) &= \sum_Y \sum_X xP(X=x, Y=y) \\ &= \sum_X x \sum_Y P(X=x, Y=y) \\ &= \sum_X xP(X=x) = E(X) \end{aligned}$$

Now for the continuous case:

$$\begin{aligned} E[E(X|Y)] = E[g(Y)] &= \int_{-\infty}^{\infty} g(y)f_Y(y)\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf_{X|Y}(x|y)f_Y(y)\,dydx \\ &= \int_{-\infty}^{\infty} x\,dx \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \\ &= \int_{-\infty}^{\infty} x\,dx[f_x(x)] = \int_{-\infty}^{\infty} xf_x(x) = E(X) \end{aligned}$$

Notice that in both of these proofs we utilized the law of total *probability*. For this reason, sometimes people just refer to the law of total expectation as the law of total probability.

Here's a "cute" example regarding the power of the law of total expectation: Suppose an insurance company receives N claims every year where N is a random variable with mean μ_n and variance σ_N^2. Now, let X_i be the size of the ith claim (in dollars) received by the insurance company so that $\sum_{i=1}^{N} X_i$ represents the total loss incurred by the insurance company throughout the year. We'll make the following assumptions about the variables X_i:

- $X_1, \ldots, X_n$ are independent and identically distributed each with mean μ_X and variance σ_X^2
- Each of the X_i's are independent of N

Compute the expected value for the total losses incurred by the insurance company throughout the year, that is, calculate:

$$E\left[\sum_{i=1}^{N} X_i\right]$$

To do this we'll use the law of total expectation:

$$E\left[\sum_{i=1}^{N} X_i\right] = E\left[E\left(\sum_{i=1}^{N} X_i|N\right)\right]$$

First, we'll consider the inner expectation. Because each of the X_i's are independent of N, we can rewrite this as:

$$E\left(\sum_{i=1}^{N} X_i|N\right) = E\left(\sum_{i=1}^{n} X_i|N=n\right) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = n\mu_X$$

Now, returning to the original expression for the expectation,

$$E\left[E\left(\sum_{i=1}^{N} X_i|N\right)\right] = E\left[N\mu_X\right] = \mu_X E(N) = \mu_X\mu_N$$

which is our final answer for the expected value. Intuitively, this result makes sense. The expected value for the total losses should be equal to the expected value of the number of claims filed times the expected value of the size of the claim.

Now, we'll calculate the variance of the total losses incurred by the insurance company:

$$\text{Var}(N) = \text{Var}\left(\sum_{i=1}^{N} X_i\right)$$

We'll start by applying the computationally friendly definition of the variance:

$$\text{Var}(N) = E\left[\left(\sum_{i=1}^{N} X_i\right)^2\right] - \left[E\left(\sum_{i=1}^{N} X_i\right)\right]^2$$

We just calculated the first moment of N so what's left to do is compute the second moment. Notice that we can rewrite what's being squared as a double sum:

$$\left(\sum_{i=1}^{N} X_i\right)^2 = \left(\sum_{i=1}^{N} X_i\right)\left(\sum_{j=1}^{N} X_j\right) = \sum_{i=1}^{N}\sum_{j=1}^{N} X_i X_j$$

Thus, our expression for the variance becomes:

$$\text{Var}(N) = E\left[\sum_{i=1}^{N}\sum_{j=1}^{N} X_i X_j\right] - \left[E\left(\sum_{i=1}^{N} X_i\right)\right]^2$$

Now, a quick detour. Let's compute the conditional expectation of the second moment, conditioning on N. Because all of the X_i's are independent of N, we come to the following:

$$E\left[\sum_{i=1}^{N}\sum_{j=1}^{N} X_i X_j|N\right] = E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j|N=n\right] = E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j\right]$$

Break this expectation up into two parts: One part being when $i = j$ and the other being when $i \neq j$:

$$\begin{aligned} E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j\right] &= E\left[\sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j\right] \\ &= E\left[\sum_{i=1}^{n} X_i^2\right] + E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j\right] \\ &= \sum_{i=1}^{n} E(X_i^2) + \sum_{i=1}^{n}\sum_{j=1}^{n} E(X_i)E(X_j) \end{aligned}$$

In the final step we utilized the linearity of the expectation as well as the fact that all of the X_i's are independent. At this point, we can substitute known values into the expression. The second moment of X_i can be founded by adding the variance and the squared mean together to afford $E(X_i^2) = \sigma_X^2 + \mu_X^2$ and the expected values of all of the X_i's are μ_X. Thus,

$$\begin{aligned}\sum_{i=1}^{n} E(X_i^2) + \sum_{i=1}^{n}\sum_{j=1}^{n} E(X_i)E(X_j) &= n(\sigma_X^2 + \mu_X) + n(n-1)\mu_X\mu_X \\ &= n\sigma_X^2 + n\mu_X^2 + n^2\mu_X^2 - n\mu_X^2 \\ &= n\sigma_X^2 + n^2\mu_X^2\end{aligned}$$

Recall that this is the final expression for the second moment of the expression conditioned on N. So, we see that

$$E\left[\sum_{i=1}^{N}\sum_{j=1}^{N} X_iX_j|N\right] = N\sigma_X^2 + N^2\mu_X^2$$

which allows us to compute the second moment (what we were originally looking for prior to the short detour) using the law of total expectation as follows:

$$\begin{aligned}E\left[\sum_{i=1}^{N}\sum_{j=1}^{N} X_iX_j\right] &= E\left[E\left(\sum_{i=1}^{N}\sum_{j=1}^{N} X_iX_j|N\right)\right] \\ &= E\left(N\sigma_X^2 + N^2\mu_X^2\right) \\ &= \sigma_X^2 E(N) + \mu_X^2 E(N^2) \\ &= \sigma_X^2(\mu_N) + \mu_X^2(\sigma_N^2 + \mu_N^2)\end{aligned}$$

This is our final answer for the second moment of N. It's been a long road but finally, we are able to compute the variance:

$$\begin{aligned}\text{Var}(N) &= E\left[\sum_{i=1}^{N}\sum_{j=1}^{N} X_iX_j\right] - \left[E\left(\sum_{i=1}^{N} X_i\right)\right]^2 \\ &= \sigma_X^2\mu_N + \mu_X^2(\sigma_N^2 + \mu_N^2) - (\mu_X\mu_N)^2 \\ &= \sigma_X^2\mu_N + \sigma_N^2\mu_X^2\end{aligned}$$

Intuitively, one may suspect that the variance for this random variable N is solely $\text{Var}(N) = \sigma_X^2\mu_N$. This is wrong.

7.4 Variance

The joint variance, similar to the joint expectation, does not change at all with jointly distributed variables. Once again, however, we can define a *conditional variance*. The conditional variance of two jointly distributed variables X and Y is given by

$$\text{Var}(X|Y) = E[(X - E[X|Y])^2|Y] = E[(X - E[X|Y = y])^2|Y = y]$$

What's nice about this expression is that the conditional variance is the variance of the conditional distribution. That is, $\text{Var}(X|Y)$ is the variance of $f_{X|Y}$ for some random variables X, Y. Similar to the conditional expectation, the conditional variance can be thought of as random variable. Here is an example to illustrate the computation involved behind the conditional variance: Suppose X, Y are

jointly continuous with a joint PDF of $f_{X,Y} = xe^{-x(1+y)}$ for $x, y > 0$. Compute $\text{Var}(X|Y)$. First, we'll need to find the conditional distribution of X given Y:

$$f_{X|Y} = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{xe^{-x(1+y)}}{f_Y(y)}$$

Quick detour to find the marginal distribution of Y:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \int_0^{\infty} xe^{-x(1+y)}\,dx$$

Notice that this is the form of a Gamma$(2, (1+y)^{-1})$ without the normalizing factor. So, we can evaluate it as follows:

$$f_Y(y) = \int_0^{\infty} xe^{-x(1+y)}\,dx = [(1+y)^{-1}]^2\Gamma(2) = (1+y)^{-2}$$

Now, returning to the conditional distribution,

$$f_{X|Y} = \frac{xe^{-x(1+y)}}{f_Y(y)} = x(1+y)^2e^{-x(1+y)}$$

Once again we have a Gamma$(2, (1+y)^{-1})$, this time with the normalizing factor. So, we can go ahead and use the known expression for the variance of the Gamma distribution:

$$\text{Var}(X|Y) = \alpha\beta^2 = (2)\left(\frac{1}{1+y}\right)^2 = \frac{2}{(1+y)^2}$$

7.4.1 Law of Total Variance

The Law of Total Variance states

$$\text{Var}(X) = \text{Var}[E(X|Y)] + E[\text{Var}(X|Y)]$$

for some random variables X and Y with finite means and variances. Here is a proof of the Law of Total Variance, beginning with the definition of the variance:

$$\begin{aligned}
\text{Var}(X) &= E\{[X - E(X)^2]\} \\
&= E\left\{[X - E(X|Y) + E(X|Y) - E(X)]^2\right\} \\
&= E\left[E\left\{[X - E(X|Y) + E(X|Y) - E(X)]^2|Y\right\}\right] \\
&= E\left[E\left\{(X - E(X|Y))^2 + (E(X|Y) - E(X))^2 + 2(E(X|Y) - E(X))(X - E(X|Y))|Y\right\}\right] \\
&= E\left[E\left\{E[(X - E(X|Y))^2|Y] + E[(E(X|Y) - E(X))^2|Y] + h(Y)E[(X - E(X|Y))|Y]\right\}\right] \\
&= E\left[\text{Var}(X|Y) + \text{Var}[E(X|Y)] + h(Y)[E(X|Y) - E(X|Y)]\right] \\
&= E[\text{Var}(X|Y)] + \text{Var}[E(X|Y)]
\end{aligned}$$

QED

Here is an example regarding an application of the Law of Total Variance: Consider the set of random variables $Z_1, Z_2 \sim \text{iid}\mathcal{N}(0,1)$ and the bivariate normal distribution

$$X = \mu_x + \sigma_x Z_1$$
$$Y = \mu_y + \sigma_y \rho Z_1 + \sigma_y \sqrt{1-\rho^2} Z_2$$

for some parameters $\mu_x, \mu_y, \sigma_x, \sigma_y$, and ρ where $-1 \leq \rho \leq 1$. First, determine the conditional distribution of $Y|X$: To do this we can substitute $X = \mu_x + \sigma_x X$ into our expression for the random variable Y:

$$Y = \mu_y + \sigma_y \rho Z_1 + \sigma_y \sqrt{1-\rho^2} Z_2 \Rightarrow Y = \mu_y + \frac{\sigma_y \rho}{\sigma_x}(X - \mu_x) + \sigma_y \sqrt{1-\rho^2} Z_2$$

This is the conditional distribution $Y|X$. Notice, $Y|X \sim \mathcal{N}(\mu_y + \frac{\sigma_y \rho}{\sigma_x}(x - \mu_x), \sigma_y^2(1-\rho^2))$. So, to compute the variance (and expectation for that matter) we can use the known expressions for the normal distribution. This is easy because the expected value and variance are staring at us:

$$E(Y|X) = \mu_y + \frac{\sigma_y \rho}{\sigma_x}(X - \mu_x)$$
$$\text{Var}(Y|X) = \sigma_y^2(1-\rho^2)$$

Now, what is the variance of Y? Unless you're brain dead or just haven't read anything else up to this point, you probably already realize we'll use the Law of Total Variance to do this:

$$\begin{aligned}
\text{Var}(X) &= \text{Var}[E(X|Y)] + E[\text{Var}(X|Y)] \\
&= \text{Var}\left[\mu_y + \frac{\sigma_y \rho}{\sigma_x}(X - \mu_x)\right] + E[\sigma_y^2(1-\rho^2)] \\
&= \left(\frac{\sigma_y^2 \rho^2}{\sigma_x^2}\right)\text{Var}(X - \mu_x) + \sigma_y^2(1-\rho^2) = \left(\frac{\sigma_y^2 \rho^2}{\sigma_x^2}\right)\text{Var}(X) + \sigma_y^2(1-\rho^2) \\
&= \left(\frac{\sigma_y^2 \rho^2}{\sigma_x^2}\right)\sigma_x^2 + \sigma_y^2(1-\rho^2) \\
&= \sigma_y^2 \rho^2 + \sigma_y^2(1-\rho^2) = \sigma_y^2
\end{aligned}$$

This result makes perfect sense! When conditioned on the "other part" of the bivariate normal, $Y|X$ has a variance of σ_y^2 which is just the variance of Y. Also, notice that:

$$\text{Var}(Y) = \sigma_y^2 \leq \rho^2 \sigma_y^2 = \text{Var}[E(Y|X)]$$

Why did I mention this? I don't know it's pretty late I haven't gone to bed before 2am in over a month I'm kinda losing it.

That being said, I mention this because notice that everything in the expression for the law of total variance is nonnegative (due to the fact that the variance is defined to be nonnegative). Thus, we have:

$$\text{Var}(X) \leq \text{Var}[E(X|Y)]$$

This is known as the Rao-Blackwell inequality and is vastly important in statistics. On the surface, it posits that the variance of a random variable can *always* be decreased by conditioning it on another random variable, which is exactly what we saw in the previous example.

7.5 Covariance

The covariance between two random variables X and Y with finite means μ_x and μ_y and finite variances σ_x^2 and σ_y^2, respectively, is given by:

$$\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

A more computationally friendly form of this expression can be derived as follows:

$$\begin{aligned} E[(X - \mu_x)(Y - \mu_y)] &= E[XY - X\mu_y - Y\mu_x + \mu_x\mu_y] \\ &= E(XY) - E[X\mu_y] - E[Y\mu_x] + E[\mu_x\mu_y] \\ &= E(XY) - \mu_y E[X] - \mu_x E[Y] + \mu_x\mu_y \\ &= E(XY) - \mu_y\mu_x - \mu_x\mu_y + \mu_x\mu_y \\ &= E(XY) - \mu_x\mu_y \end{aligned}$$

Thus, we can also express the covariance as

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

and when X and Y are independent,

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0$$

which relates to *correlation*, something that I'll talk about in a minute or two.

Some properties of the covariance that are easy to proof and left as an exercise for the (absolutely incredible btw) reader:

1. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
2. $\text{Cov}(X, X) = \text{Var}(X)$
3. $\text{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \text{Cov}(X_i, Y_j)$

A result that can be pseudo-derived from facts 2 and 3 is that

$$\text{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j X_j \right) = \sum_{i=1}^{n} a_i \text{Var}(X_i) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \text{Cov}(X_i, X_j)$$

Hmmm... In order to compute the double sum of the covariances, let's rewrite it as an $n \times n$ matrix as follows:

$$\begin{bmatrix} a_1^2\text{Cov}(X_1, X_1) & a_1 a_2 \text{Cov}(X_1, X_2) & \dots & a_1 a_n \text{Cov}(X_1, X_n) \\ a_2 a_1 \text{Cov}(X_2, X_1) & a_2^2 \text{Cov}(X_2, X_2) & \dots & a_1 a_n \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ a_n a_1 \text{Cov}(X_b, X_1) & a_n a_2 \text{Cov}(X_n, X_2) & \dots & a_n^2 \text{Cov}(X_n, X_n) \end{bmatrix}$$

On the main diagonal, we have

$$\sum_{i=1}^{n} a_i^2 \text{Cov}(X_i, X_i) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i)$$

and on the off-diagonal entries we can make use of the symmetry of the matrix to afford the expression

$$2\sum_{i<j}^{n}\sum_{j=1}^{m} a_i a_j \text{Cov}(X_i, X_i)$$

for the sum of the rest of the entries in the matrix. With all of that, we come to the powerful expression:

$$\sum_{i=1}^{n} a_i \text{Var}(X_i) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \text{Cov}(X_i, X_j) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i) + 2\sum_{i<j}^{n}\sum_{j=1}^{m} a_i a_j \text{Cov}(X_i, X_i)$$

Moreover, if $X_1, \ldots, X_n$ are mutually independent,

$$\sum_{i=1}^{n} \text{Var}(a_i X_i) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i)$$

which agrees with our intuition about the linearity of the variance.

7.5.1 Correlation

The correlation between two random variables X and Y with respective finite variances σ_x^2 and σ_y^2 is given by:

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}, \quad \text{for } -1 \leq \rho \leq 1$$

Physically, the correlation measures the *strength* of the linear association between the random variables. The closer the correlation is to ± 1, the stronger the association. An important aspect of the correlation to mention is that *uncorrelated* random variables (when $\rho = 0$) does not imply that the random variables are independent. They can be, but this is not always the case. Here's an example: Let X be a discrete random variable with PMF

$$P(X = -1) = P(X = 1) = \frac{1}{4}, P(X = 0) = \frac{1}{2}$$

and let $Y = X^2$. Show that X and Y are uncorrelated yet are dependent on eachother. Clearly, X and Y are not independent. Why is this clear? Well, given that $Y = X^2$, one can imagine that the value of Y will certainly be influenced by the value of X. Thus, these random variables are not independent. Now, to show that they are uncorrelated. Using the definition of the correlation we see that in order for $\rho = 0$ we need $\text{Cov}(X, Y) = 0$. So, let's begin by computing the covariance:

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = E(X^3) - E(X)E(X^2)$$

It appears we will need the first three moments of X. Since the PMF of X involves only integers on $[-1, 1]$, we can do this by enumeration pretty easily. First, the first moment:

$$\begin{aligned} E(X) &= \sum_{X} xP(X = x) \\ &= (1)P(X = 1) + (-1)P(X = -1) + (0)P(X = 0) \\ &= \frac{1}{4} - \frac{1}{4} = 0 \end{aligned}$$

Next, the other two moments:

$$\begin{aligned}
E(X^2) &= \sum_X x^2 P(X^2 = x) \\
&= (1)^2 P(X^2 = 1) + (0)^2 P(X^2 = 0) \\
&= (1)^2 [P(X = 1) + P(X = -1)] \\
&= \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \\
E(X^3) &= \sum_X x^3 P(X^3 = x) \\
&= (1)^3 P(X^3 = 1) + (-1)^3 P(X^3 = -1) + (0) P(X^3 = 0) \\
&= \frac{1}{4} - \frac{1}{4} = 0
\end{aligned}$$

Thus, the covariance becomes

$$\mathrm{Cov}(X, Y) = E(X^3) - E(X)E(X^2) = 0 - 0 \cdot \frac{1}{2} = 0$$

and the random variables X and Y are uncorrelated.

8 Random Variable Transformations

Random variable transforms allow us to relate probability distributions to one another in a concise manner. For example, consider two random variables $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ with a *correlation coefficient* of $-1 \leq \rho \leq 1$. The joint PDF of X and Y is said to have a *bivariate normal distribution* with the following PDF:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{ \frac{-1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_x}{\sigma_x} \right)^2 - 2\rho \left(\frac{x-\mu_x}{\sigma_x} \right) \left(\frac{y-\mu_y}{\sigma_y} \right) + \left(\frac{y-\mu_y}{\sigma_y} \right)^2 \right] \right\}$$

I'm gonna be honest I had to get up a take a quick snack break after writing that just now. How does anyone expect to write out that distribution every time we wanna express the bivariate normal distribution? They shouldn't, and they can't! So, instead of having to write this out every time, we can express this distribution using variable transforms. If we are given $Z_1, Z_2 \sim \mathcal{N}(0,1)$ and parameters $\mu_1, \mu_2, \sigma_1, \sigma_2$, and $\rho \in [-1,1]$, and we define,

$$X_1 = \mu_1 + \sigma_1 Z_1$$
$$X_2 = \mu_2 + \sigma_2 \rho Z_2 + \sigma_2 \sqrt{1-\rho^2} Z_2$$

then the joint distribution of X_1 and X_2 is said to have the bivariate normal distribution.

Another example of a case where we typically express a distribution in terms of a variable transform instead of a PDF is Fisher's F-distribution (not Fisher Gandel, sadly. It's some other Fisher). A continuous random variable U is said to have an F-distribution if and only if there exist independent random variables $X \sim \chi_n^2$ and $Y \sim \chi_m^2$ such that

$$U = \frac{X/n}{Y/m}$$

i.e., the F-distribution is defined as the ratio of independent chi-square distributions to their degrees of freedom. Finding this PDF is by no means challenging, some might consider it easy to do so. Either way, the F-distribution is more often than not referred to as a variable transform as opposed to its PDF because it is far easier to recognize and work with. With these two examples as motivation, we will now look at a few methods of going about transforming random variables.

8.1 Method of CDFs

Suppose $X \sim f_X(x)$ and $Y = g(X)$ and we are interested in the PDF of Y. The steps we take to find the PDF of Y using the CDF method is as follows:

1. Identify the support of Y. To do this, we'll first need to find the support of X.
2. For an arbitrary $y \in \text{supp}(Y)$, compute the CDF of Y, $F_Y(y) = P(Y \leq y)$.
3. Compute $P(Y \leq y) = P(g(X) \leq y)$, treating the random variable X in terms of y.
4. Compute $f_Y(y) = F_Y'(y)$ using the fundamental theorem of calculus.

Let's do an example to illustrate the process. Consider a random variable $Z \sim \mathcal{N}(0,1)$. Evaluate the PDF of $Y = Z^2$.

Step 1: The PDF of the standard normal distribution is

$$\varphi(z) = \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}} \quad \text{for } -\infty < z < \infty$$

Thus, the support of Z is $(-\infty, \infty)$. Since $Y = Z^2$, the support of Y becomes $(0, \infty)$ because Y does not take on negative values.

Step 2: For a $y \in (0, \infty)$, we have that any $y > 0$ will have $F_Y(y) = P(Y \leq y)$, otherwise $F_Y(y) = 0$.

Step 3: Now, we'll find the CDF of Y as follows:

$$F_Y(y) = P(Y \leq y) = P(Z^2 \leq y) = P(|Z| \leq \sqrt{y}) = P(-\sqrt{y} \leq Z \leq \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y})$$

where Φ is the CDF of Z.

Step 4: What's left to do is evaluate the PDF of Y using the fundamental theorem of calculus:

$$\begin{aligned}
f_Y(y) &= \frac{d}{dy}\left(\Phi(\sqrt{y}) - \Phi(-\sqrt{y})\right) = \frac{d}{dy}\int_{-\sqrt{y}}^{\sqrt{y}} \varphi(z)\, dz \\
&= \left[\frac{d}{dy}\int_{-\infty}^{\sqrt{y}} \varphi(z)\, dz\right] - \left[\frac{d}{dy}\int_{-\infty}^{-\sqrt{y}} \varphi(z)\, dz\right] \\
&= \varphi(\sqrt{y})\left[\frac{d}{dy}(\sqrt{y})\right] - \varphi(-\sqrt{y})\left[\frac{d}{dy}(-\sqrt{y})\right] \\
&= \frac{1}{2\sqrt{y}}\left(\frac{e^{-\frac{1}{2}(\sqrt{y})^2}}{\sqrt{2\pi}}\right) + \frac{1}{2\sqrt{y}}\left(\frac{e^{-\frac{1}{2}(-\sqrt{y})^2}}{\sqrt{2\pi}}\right) \\
&= \frac{e^{-y/2}}{\sqrt{2\pi y}} = \frac{y^{(\frac{1}{2}-1)}e^{-y/2}}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})}
\end{aligned}$$

Thus, we see that $Y \sim \text{Gamma}(\frac{1}{2}, 2)$ with a PDF given by

$$f_Y(y) = \frac{e^{-y/2}}{\sqrt{2\pi y}} \quad \text{for } y > 0$$

Notice, we've just shown that the standard normal distribution-squared is a gamma distribution. Furthermore, the standard normal distribution-squared is a chi-square distribution with one degree of freedom.

Here is another example: Suppose $X, Y \sim \text{iid}\mathcal{U}(0, 1)$. Show that the PDF of $S = X + Y$ follows a triangular distribution.

Step 1: The supports of both X and Y are $[0, 1]$. Thus, the support of S is $[0, 2]$.

Step 2: Now, we'll compute the CDF of S:

$$F_S(s) = P(S \leq s) = P(X + Y \leq s) = \iint_{X,Y} f_{X,Y}(x, y)\, dxdy$$

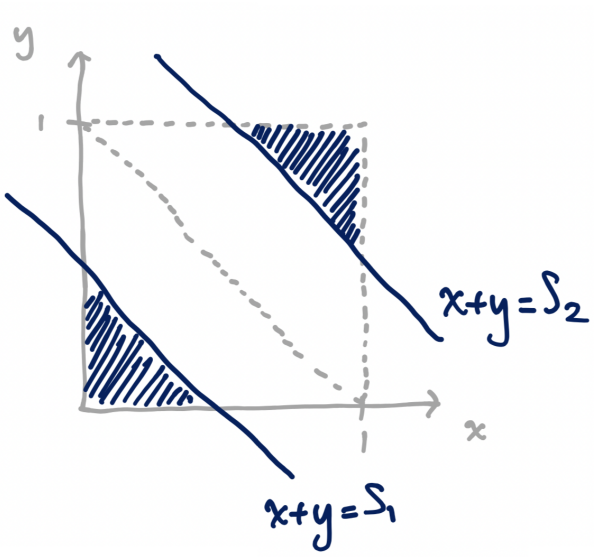
Ah! We forgot to find the joint PDF of X and Y. No matter, they're pretty simple anyway:

$$f_X(x) = \frac{1}{1-0} = 1 = f_Y(y)$$

Since X and Y are independent, $f_{X,Y}(x,y) = f_X(x)f_Y(y) = 1$. Back to computing the CDF:

$$\iint_{X,Y} f_{X,Y}(x,y)\,dxdy = \iint_{X,Y} dxdy$$

The bounds on S are not intuitive: Since we are integrating over the $[0,1] \times [0,1]$ region in the



XY-plane we're encouraged to split up the integration into two parts, one where $0 \le s_1 < 1$ and the other where $1 \le s_2 < 2$. The bounds on both regions are illustrated below:

S_1	S_2
$0 \le x \le s_1 - y$	$s_2 - y \le x \le 1$
$0 \le y \le s_1$	$s_2 - 1 \le y \le 1$

Now we can evaluate the CDF of S:

$$\begin{aligned}
\iint_{X,Y} f_{X,Y}(x,y)\,dxdy &= \int_0^{s_1}\int_0^{s_1-y} dxdy \\
&= \int_0^{s_1} (s_1 - y)\,dy \\
&= \left[s_1 y - \frac{1}{2}y^2\right]_0^{s_1} \\
&= s_1^2 - \frac{1}{2}s_1^2 = \frac{1}{2}s^2 = F_S(s)
\end{aligned}$$

By the fundamental theorem of calculus, we have that

$$f_S(s) = \frac{d}{ds}F_S(s) = \frac{d}{ds}\left[\frac{1}{2}s^2\right] = s \quad \text{for } 0 \le s < 1$$

Now to evaluate the CDF in the region restricted by S_2. Notice that the shaded region is actually the *complement* of the CDF and so we'll be interested in $1 - F_S(s)$ when everything is said and done:

$$\begin{aligned}
F_S(s) = P(S \leq s) = P(X + Y \leq s) &= \iint_{X,Y} f_{X,Y}(x, y)\, dxdy \\
&= \int_{s_2-1}^{1} \int_{s_2-y}^{1} dxdy \\
&= \int_{s_2-1}^{1} (1 - s_2 + y)\, dy \\
&= \left[y - s_2 y + \frac{1}{2} y^2 \right]_{s_2-1}^{1} \\
&= \left(1 - s_2 + \frac{1}{2} \right) - \left((s_2 - 1) - s_2(s_2 - 1) + \frac{1}{2}(s_2 - 1)^2 \right) \\
&= 1 - s_2 + \frac{1}{2} - s_2 + 1 + s_2^2 - s_2 - \frac{1}{2}(s_2^2 - 2s_2 + 1) \\
&= s_2^2 - 3s_2 + \frac{5}{2} - \frac{1}{2} s_2^2 + s_2 - \frac{1}{2} \\
&= \frac{1}{2} s^2 - 2s_2 + 2 = F_S(s)
\end{aligned}$$

Recall that we are interested in $1 - F_S(s)$. So, by the fundamental theorem of calculus, we have that

$$f_S(s) = \frac{d}{ds}\left[1 - F_S(s) \right] = \frac{d}{ds}\left[1 - \frac{1}{2} s^2 + 2s_2 - 2 \right] = 2 - s \quad \text{for } 1 \leq s < 2$$

Thus, the PDF of $S = X + Y$ is:

$$f_S(s) = \begin{cases} s & 0 \leq s < 1 \\ 2 - s & 1 \leq s < 2 \\ 0 & \text{otherwise} \end{cases}$$

Here is yet another example! This time, we'll find the PDF of the $Y = e^X$ and $X \sim \mathcal{N}(\mu, \sigma^2)$ so Y is log-normally distributed.

Step 1: Since X is normally distributed it is supported on the entire real number line. So, given that $Y = e^X$, Y will be supported by the positive reals. Thus, $Y > 0$.

Step 2/3: The CDF of Y will be given by

$$F_Y(y) = P(Y \leq y) = P(e^X \leq y) = P(X \leq \ln y) = F_X(\ln y)$$

and the CDF of X is given by

$$F_X(x) = \int_{-\infty}^{x} f_X(t) dt$$

where $f_X(x)$ is just the PMF of the normal distribution, and therefore the CDF of Y is given by

$$F_Y(y) = F_X(\ln y) = \int_{-\infty}^{\ln y} f_X(t) dt$$

Step 4: Evaluate the PDF of Y using the fundamental theorem of calculus and the Leibniz rule:

$$f_Y(y) = F_Y'(y) = f_X(\ln y)\frac{d}{dy}\Big[\ln y\Big] = \frac{f_X(\ln y)}{y}$$

Thus, using the PMF of the normal distribution, we find that the PMF of Y is given by:

$$f_Y(y) = \frac{1}{y\sigma\sqrt{2\pi}}\left(e^{-\frac{1}{2}(\frac{\ln y-\mu}{\sigma})^2}\right) \quad \text{for } 0 < y < \infty$$

8.2 Method of Convolutions

8.2.1 Convolutions

A convolution is a binary operation on two functions that expresses how the shape of one is modified by the other. It is defined as the integral of the product of the two functions after one is reversed and shifted. For two functions f and g, their convolution is given by

$$(f * g)(t) = \int_{-\infty}^{\infty} f(s)g(t-s)\,ds = \int_{-\infty}^{\infty} f(t-s)g(s)\,ds = (g * f)(t)$$

where we say that *f is convolved with g at t*. That's the technical definition, at least. In regards to probability, convolutions offer a shortcut to finding the sum of two jointly distributed independent random variables. I won't talk about it here, but convolutions have applications in all sorts of fields, in particular with Fourier and Laplace transforms (I mention these only because I think they're really cool).

8.2.2 Within Probability

As mentioned briefly above, the method of convolutions is only applicable for *jointly distributed independent random variables*. With that in mind, suppose X and Y are two jointly distributed independent random variables with joint PDF $f_{X,Y}(x, y)$. Furthermore, suppose that we are interested in the sum of these variables, denoted $U = X + Y$. Using the CDF method, we can compute the PDF of U as follows:

$$\begin{aligned} F_U(u) = P(U \le u) &= P(X + Y \le u) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{u-x} f_{X,Y}(x, y)\,dydx \end{aligned}$$

Then, the PDF of U, denoted $f_U(u)$ is given by:

$$\begin{aligned} f_U(u) &= \frac{d}{du}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{u-x} f_{X,Y}(x, y)\,dydx\right] \\ &= \int_{-\infty}^{\infty} f_{X,Y}(x, u-x)\,dx \end{aligned}$$

If we had chosen constants limits of integration for y instead, we would have arrived at the equivalent result,

$$f_U(u) = \int_{-\infty}^{\infty} f_{X,Y}(u-y, y)\,dy$$

Now, since we assumed that X and Y were independent, we can express their joint PDF as $f_{X,Y}(x, y) = f_X(x)f_Y(y)$, affording the integral

$$\int_{-\infty}^{\infty} f(x, u-x)\,dx = \int_{-\infty}^{\infty} f_X(x)f_Y(u-x)\,dx = (f_X * f_Y)(u)$$

Thus, we've shown that the sum of two jointly distributed independent random variables is given by their convolution, that is,

$$f_S(s) = f_{X+Y}(x+y) = (f_X * f_Y)(s)$$

Here's an example utilizing the convolution formula to find a PDF: Suppose X and Y are independent and identically distribution $\mathcal{N}(0, \frac{1}{2})$ random variables and let $S = X + Y$. Find the PDF of S. Before we can convolve X and Y we need their PDFs:

$$f_X(x) = \frac{e^{-\frac{1}{2}(\sqrt{2}x)^2}}{\sqrt{\frac{1}{2}}\sqrt{2\pi}} = \frac{e^{-x^2}}{\sqrt{\pi}} = f_Y(y)$$

We're off to a great start. Now to convolve the random variables:

$$f_S(s) = (f_X * f_Y)(s) = \int_{-\infty}^{\infty} f_{X,Y}(x, s-x)\, dx = \int_{-\infty}^{\infty} f_X(x) f_Y(s-x)\, dx$$

Great! Note that we can separate the joint PDF into the product of the the PDFs for X and Y because the random variables are independent. Now, we can substitute the PDFs for X and Y and evaluate the integral:

$$\begin{aligned}
\int_{-\infty}^{\infty} f_X(x) f_Y(s-x)\, dx &= \int_{-\infty}^{\infty} \left(\frac{e^{-x^2}}{\sqrt{\pi}}\right)\left(\frac{e^{-(s-x)^2}}{\sqrt{\pi}}\right) dx = \frac{1}{\pi}\int_{-\infty}^{\infty} e^{-x^2-(s-x)^2}\, dx \\
&= \frac{1}{\pi}\int_{-\infty}^{\infty} e^{-x^2-s^2-x^2+2sx}\, dx \\
&= \frac{e^{-s^2}}{\pi}\int_{-\infty}^{\infty} e^{-2(x^2-sx+\frac{1}{4}s^2-\frac{1}{4}s^2)}\, dx \\
&= \frac{e^{-s^2}}{\pi}\int_{-\infty}^{\infty} e^{-2(x^2-\frac{1}{2}s)-\frac{1}{4}s^2}\, dx \\
&= \frac{e^{-\frac{s^2}{2}}}{\pi}\int_{-\infty}^{\infty} e^{-2(x^2-\frac{1}{2}s)}\, dx
\end{aligned}$$

Here we're left with an integral of the form $\mathcal{N}(\frac{s}{2}, \frac{1}{2})$ without a normalizing factor and therefore the integral will evaluate to that normalizing factor. Thus,

$$\frac{e^{-\frac{s^2}{2}}}{\pi}\int_{-\infty}^{\infty} e^{-2(x^2-\frac{1}{2}s)}\, dx = \frac{e^{-\frac{s^2}{2}}}{\pi}\left(\sqrt{\frac{1}{2}}\sqrt{\pi}\right) = \frac{e^{-\frac{s^2}{2}}}{\sqrt{2\pi}}$$

8.3 Method of Jacobians

Before getting into the more general method of Jacobians, involving multivariables, we'll consider the case of a single variable transformation. Suppose we have a random variable X and another random variable $Y = g(X)$ for some function $g : \mathbb{R} \to \mathbb{R}$ which is strictly monotone. If X has a PDF $f_X(x)$, then the PDF of Y is given by

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{d}{dy}\left[g^{-1}(y)\right]\right|$$

Here is an example to begin to illustrate the process of using this formula to find the PDF of the transformed variable, in this case Y: Suppose we have two random variables, $\Theta \sim \mathcal{U}(-\frac{\pi}{2}, \frac{\pi}{2})$ and

$X = \alpha + \gamma \tan\Theta$ where $\alpha, \gamma \in \mathbb{R}$ and $\gamma > 0$. Note that $\tan\theta$ is monotone increasing, so, we are okay applying the formula above:

$$f_\Theta = \frac{1}{\frac{\pi}{2} - (-\frac{\pi}{2})} = \frac{1}{\pi}$$

$$X = \alpha + \gamma \tan\Theta \Rightarrow \Theta = \arctan\left(\frac{X-\alpha}{\gamma}\right) = g^{-1}(X)$$

$$\frac{d}{dX} g^{-1}(X) = \frac{d}{dX}\left[\arctan\left(\frac{X-\alpha}{\gamma}\right)\right] = \frac{1}{1 + (\frac{X-\alpha}{\gamma})^2} \cdot \frac{1}{\gamma}$$

Thus, applying the formula, the PDF of X becomes

$$f_X = f_\Theta(g^{-1}(X)) \cdot \frac{1}{1 + (\frac{X-\alpha}{\gamma})^2} \cdot \frac{1}{\gamma} = \frac{1}{\gamma\pi}\left(\frac{1}{1 + (\frac{X-\alpha}{\gamma})^2}\right)$$

which is a Cauchy(α, γ) distribution.

Here is another example that I'll include which uses this formula because I'm studying for my midterm tomorrow and I'm gonna do it as practice: Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = e^X$ so Y is log-normally distributed. Let's find the PDF of Y. We've identified the PDF of X so many times now I don't wanna write it out again, so I'll just denote it as $f_X(x)$. The inverse transformation for Y is

$$X = g^{-1}(Y) = \ln Y$$

and therefore the derivative with respect to y of $g^{-1}(y)$ is given by:

$$\frac{d}{dy}\left[\ln y\right] = \frac{1}{y}$$

Also, note that while X is supported on all reals because it is normally distributed, Y is supported by the positive reals. So, the PDF of Y becomes

$$\begin{aligned} f_Y(y) &= f_X(\ln y) \cdot \frac{1}{y} \\ &= \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\ln y - \mu}{\sigma})^2} \quad \text{for } y > 0 \end{aligned}$$

Now, the case of multiple variables. Suppose that X and Y are jointly distributed random variables and we are interested in the PDF of a random variable $U = g(X, Y)$. The method of Jacobians will offer us a general method of finding the PDF of U, as long as the transformation from X, Y to U is *injectve*, that is, the inverse transformation exists. Oftentimes, when we use this method we are looking at a system of transformations with $U = g_1(X, Y)$ and some other random variable $V = g_2(X, Y)$ (not electric potential). The general method of Jacobians (for two random variables) goes as follows:

1. First, we need the joint PDF of X and Y as well as the support of the PDF. Using set notation, the support is $\text{supp}(f_{X,Y}) = \{(x, y) : f(x, y) > 0\}$.
2. Next, we'll need to solve for X and Y in terms of U and V i.e., find the inverse transformation. We'll refer to these $X = h_1(U, V)$ and $Y = h_2(U, V)$.
3. Now, find the support of U and V. Using set notation again, this will look like $\text{supp}(f_{U,V}) = \{(u, v) : f(h_1(u, v), h_2(u, v)) \in \text{supp}(f_{X,Y})\}$.
4. Compute the Jacobian determinant of this transformation! Recall, the Jacobian determinant is given by
$$|J| = \left|\frac{\partial(x, y)}{\partial(u, v)}\right| = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}$$
where we are transforming from the xy-plane to the uv-plane.
5. The joint PDF of U and V is then given by
$$f_{U,V}(u, v) = f_{X,Y}(h_1(u, v), h_2(u, v)) \cdot |J|$$

Let's do an example to illustrate the process. Consider two independent jointly distributed random variables, both of which have an exponential distribution with parameter 1. That is, $X, Y \sim \text{Exp}(1)$. Find the joint PDF of $U = X + Y$ and $V = Y$.

Step 1: The joint PDF of X and Y will be given by the product of their individual PDFs because the random variables are independent:

$$f_{X,Y}(x, y) = e^{-x}e^{-y} = e^{-(x+y)} \quad \text{for } x, y > 0$$

Step 2: Now to find the inverse transformations:

$$U = X + Y \Rightarrow X = U - Y \Rightarrow X = U - V$$
$$V = Y \Rightarrow Y = V$$

So, we'll define the following transformations:

$$h_1(U, V) = U - V$$
$$h_2(U, V) = V$$

Step 3: To find the support of U and V we'll use the supports of X and Y:

$$X > 0 \Rightarrow U - V > 0 \Rightarrow U > V$$
$$Y > 0 \Rightarrow V > 0$$

Thus, we have that $U > V$ and $V > 0$.

Step 4: Computing this Jacobian determinant is child's play:

$$|J| = \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix} = 1$$

Step 5: All that's left to do is plug in to find the joint PDF of U and V:

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}(h_1(u,v), h_2(u,v)) \cdot |J| = f_{X,Y}(u-v, v)(1) \\ &= e^{-(u-v+v)} = e^{-u} \end{aligned}$$

Thus, the joint PDF of U and V is given by

$$f_{U,V}(u,v) = e^{-u} \quad \text{for } 0 < v < u < \infty$$

What if our variable transform had been $U = X + Y$ and $V = X - Y$? Well, Sam, that would be a whole different problem. Let's do it!

Step 1: We already showed the joint PDF for X and Y a second ago, but here it is again for reference:

$$f_{X,Y}(x,y) = e^{-(x+y)} \quad \text{for } x, y > 0$$

Step 2: The inverse transformations will require slightly more work:

$$U = X + Y \Rightarrow X = U - Y \Rightarrow V + Y = U - Y \Rightarrow Y = \frac{1}{2}(U - V)$$
$$V = X - Y \Rightarrow X = \frac{1}{2}(U + V)$$

So, we'll define the following transformations:

$$h_1(U,V) = \frac{1}{2}(U+V)$$
$$h_2(U,V) = \frac{1}{2}(U-V)$$

Step 3: Similarly, to find the support of U and V we'll use the supports of X and Y:

$$X > 0 \Rightarrow \frac{1}{2}(U+V) > 0 \Rightarrow U + V > 0 \Rightarrow V > -U$$
$$Y > 0 \Rightarrow \frac{1}{2}(U-V) > 0 \Rightarrow U - V > 0 \Rightarrow U > V$$

Thus, we have that $U > 0$ and $-U < V < U$.

Step 4: I'd still consider this Jacobian child's play:

$$|J| = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = \frac{1}{2}$$

Step 5: The new and improved joint PDF becomes

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}(h_1(u,v), h_2(u,v)) \cdot |J| = f_{X,Y}\left(\frac{1}{2}(u+v), \frac{1}{2}(u-v)\right)\left(\frac{1}{2}\right) \\ &= \frac{1}{2}e^{-u} \end{aligned}$$

Thus, our new and improved joint PDF is given by

$$f_{U,V}(u,v) = \frac{1}{2}e^{-u} \quad \text{for } 0 < u, -u < v < u$$

Now, let's compute the marginal distributions of U and V. Since the bounds on U already has bounds which are independent of V, we'll compute $f_U(u)$ first:

$$\begin{aligned} f_U(u) &= \int_{-u}^{u} \frac{1}{2} e^{-u}\, dv \\ &= \left[\frac{1}{2} v e^{-u}\right]_{-u}^{u} \\ &= u e^{-u} \quad \text{for } u > 0 \end{aligned}$$

Thus, we see that $U \sim \text{Gamma}(2, 1)$. Now the marginal of V. We'll have to switch around the bounds on our variables a little bit but eventually we find that if $v > 0$, $v < u < \infty$ and if $v < 0$, $-v < u < \infty$ (Originally, V was bounded on $-u < v < u$. Look at the graphs of $v = u$ and $v = -u$ to make sense of why we need to break this integration up into cases). First, the case where $v > 0$:

$$\begin{aligned} f_V(v) &= \int_{v}^{\infty} \frac{1}{2} e^{-u}\, du \\ &= \left[-\frac{1}{2} e^{-u}\right]_{v}^{\infty} \\ &= \frac{1}{2} e^{-v} \quad \text{for } v > 0 \end{aligned}$$

Now the case of when $v < 0$:

$$\begin{aligned} f_V(v) &= \int_{-v}^{\infty} \frac{1}{2} e^{-u}\, du \\ &= \left[-\frac{1}{2} e^{-u}\right]_{-v}^{\infty} \\ &= \frac{1}{2} e^{v} \quad \text{for } v < 0 \end{aligned}$$

Thus, $V \sim \mathcal{L}(1)$ and the PDF of V is given by:

$$f_V(v) = \frac{1}{2} e^{-|v|} \quad \text{for } -\infty < v < \infty$$

(This is known as the *double exponential* or *Laplace* distribution).

Here's another example which is a bit more involved: Consider the random variable $Z \sim \mathcal{N}(0, 1)$ and $X \sim \chi^2_m$ and suppose that Z and X are independent. The random variable

$$T = \frac{Z}{\sqrt{\frac{X}{m}}}$$

has the *Student's t-distribution* (or simply just a t-distribution). Find the PDF of T using the method of Jacobians.

Step 1: Since these random variables are independent we can find their joint PDF by taking the product of their individual PDFs:

$$f_{X,Z}(x, z) = \left(\frac{x^{\frac{m}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{m}{2}} \Gamma(\frac{m}{2})}\right)\left(\frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}}\right) = \frac{x^{\frac{m}{2}-1} e^{-\frac{1}{2}(x+z^2)}}{2^{\frac{m+1}{2}} \Gamma(\frac{m}{2})\sqrt{\pi}} \quad \text{for } x > 0, -\infty < z < \infty$$

Step 2: Now to find the inverse transformation. To do this, we'll "pair" T with $U = X$. Then,

$$T = \frac{Z}{\sqrt{\frac{X}{m}}} \Rightarrow T\sqrt{\frac{X}{m}} = Z \Rightarrow Z = T\sqrt{\frac{U}{m}}$$

$$U = X \Rightarrow X = U$$

Step 3: To find the support of T and U we'll use the supports of X and Z:

$$X > 0 \Rightarrow U > 0$$

$$-\infty < Z < \infty \Rightarrow -\infty < T\sqrt{\frac{U}{m}} < \infty \Rightarrow 0 < UT^2 < \infty$$

Since U is supported on the positive real numbers, T^2 must be positive as well which occurs for any value of T. Thus, we see that the support of T is $T > 0$.

Step 4: Computing the Jacobian determinant isn't all that bad:

$$|J| = \begin{vmatrix} \sqrt{\frac{U}{m}} & \frac{T}{2\sqrt{mU}} \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{U}{m}}$$

Step 5: Now we can plug in our results to find the joint PDF of T and U:

$$f_{T,U}(t,u) = \frac{u^{\frac{m}{2}-1}e^{-\frac{1}{2}(u+t^2u/m)}}{2^{\frac{m+1}{2}}\Gamma(\frac{m}{2})\sqrt{\pi}} \cdot \sqrt{\frac{u}{m}} = \frac{u^{\frac{m}{2}+\frac{1}{2}-1}e^{-\frac{1}{2}u(1+t^2/m)}}{2^{\frac{m+1}{2}}\Gamma(\frac{m}{2})\sqrt{m\pi}}$$

Recall that we were originally interested in the PDF of T. So, we'll find the marginal of the joint PDF above. This can be done by just integrating out the contribution from u:

$$\begin{aligned} f_T(t) &= \frac{1}{2^{\frac{m+1}{2}}\Gamma(\frac{m}{2})\sqrt{m\pi}} \int_0^\infty u^{\frac{m}{2}+\frac{1}{2}-1}e^{-\frac{u}{2}(1+\frac{t^2}{m})}\,du \\ &= \frac{1}{2^{-(\frac{m+1}{2})}2^{\frac{m+1}{2}}\Gamma(\frac{m}{2})\sqrt{m\pi}} \left[\left(\frac{1}{2}\left(1+\frac{t^2}{m}\right)\right)^{-\frac{m+1}{2}}\Gamma\left(\frac{m+1}{2}\right)\right] \\ &= \left(\frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2})\sqrt{m\pi}}\right)\left(1+\frac{t^2}{m}\right)^{-\frac{m+1}{2}} \end{aligned}$$

Oftentimes in order to carry out this integration people will make the substitution $y = (1+\frac{t^2}{m})\frac{x}{2}$ to simplify the calculation. I've never been a big fan of substitutions though because it's just making more work for yourself. I thought I'd mention it either way. Thus, the PDF of the t-distribution with m degrees of freedom is given by:

$$f_T(t) = \frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2})\sqrt{m\pi}}\left(1+\frac{t^2}{m}\right)^{-\frac{m+1}{2}} \quad \text{for } t > 0$$

8.4 Method of Moment Generating Functions

Using moment generating functions to evaluate the transformation of a random variable relies on the fact that the MGF uniquely defines a probability distribution (when it exists). So, if we can show that a transformed variable has the MGF of a known distribution, that transformed variable must follow that probability distribution. Consider the example done in the section on convolutions: Suppose $X, Y \sim \text{iid}\mathcal{N}(0, \frac{1}{2})$ and let $S = X + Y$. Find the PDF of S. The MGF of the normal distribution is known, so we can substitute the parameters μ and σ which are given to us to find the MGFs of X and Y:

$$M_X(\theta) = e^{\frac{\theta^2}{4}} = M_Y(\theta)$$

Now, we can evaluate the MGF of S:

$$\begin{aligned} M_S(\theta) &= E(e^{\theta S}) = E(e^{\theta(X+Y)}) = E(e^{\theta X}e^{\theta Y}) = E(e^{\theta X})E(e^{\theta Y}) \\ &= M_X(\theta)M_Y(\theta) = (e^{\frac{\theta^2}{4}})(e^{\frac{\theta^2}{4}}) = e^{\frac{\theta^2}{2}} \end{aligned}$$

Since $M_S(\theta) = e^{\frac{\theta^2}{2}}$ is the MGF of a standard normal distribution, it follows that $S \sim \mathcal{N}(0, 1)$.

9 Order Statistics

Order statistics are, in conjunction with rank statistics, the most fundamental tools in non-parametric statistics and inference. While *order statistics* sounds like a really scary concept, it only sort of is. As the name implies, the order statistics are simply an ordered set of random variables.

In order to discuss order statistics we first have to define a set of *independent and identically distributed* random variables (often referred to as iid random variables). Suppose $X_1, X_2, \dots, X_n$ are independent and identically distributed random variables. We can express this more concisely by writing,

$$X_1, X_2, \dots, X_n \sim \text{iid} f$$

where f is the PDF of the random variables. Now, define the following "new" random variables:

$$\begin{aligned} Y_1 &= X_{(1)} = \text{the smallest among } X_1, X_2, \dots, X_n \\ Y_2 &= X_{(2)} = \text{the second smallest among } X_1, X_2, \dots, X_n \\ &\vdots \\ Y_n &= X_{(n)} = \text{the largest among } X_1, X_2, \dots, X_n \end{aligned}$$

$Y_1, Y_2, \dots, Y_n$ are referred to as the *order statistics* of $X_1, X_2, \dots, X_n$. Graphically, the order statistics can be expressed as follows:

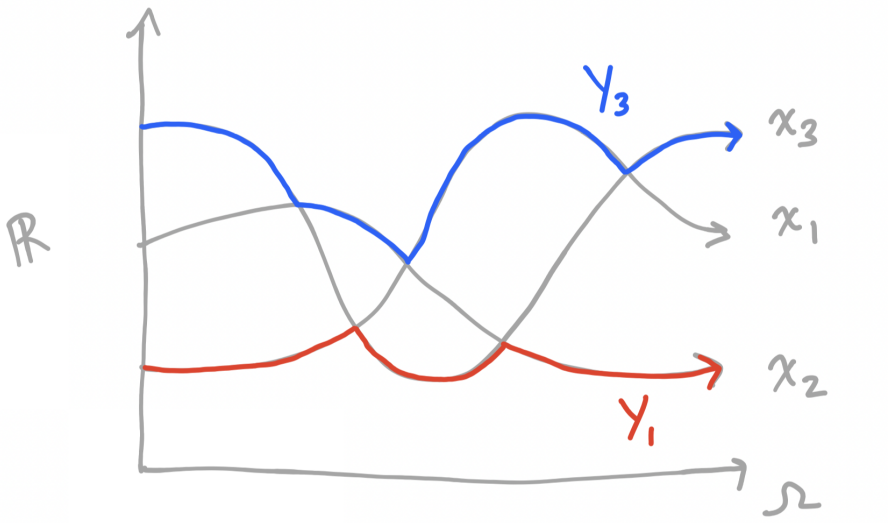


Figure 1: I drew this!

To motivate our discussion, let's first consider Y_n, the *maximum* order statistic. Since Y_n is the maximum in the set, it is given by:

$$Y_n = X_{(n)} = \max\{X_1, X_2, \dots, X_n\}$$

To find the PDF of Y_n, we'll use the CDF method. The CDF of Y_n is given by

$$\begin{aligned} P(Y_n \leq y) &= P(\max\{X_1, X_2, \dots, X_n\} \leq y) \\ &= P(X_1 \leq y, X_2 \leq y, \dots, X_n \leq y) \\ &= P(X_1 \leq y)P(X_2 \leq y) \dots P(X_n \leq y) \\ &= F_{X_1}(y) \cdot F_{X_2}(y) \cdots F_{X_n}(y) = [F_X(y)]^n \end{aligned}$$

The last substitution we made was possible by the fact that all of the random variables $X_1, \ldots, X_n$ are identically distributed. Now, we can use the fundamental theorem of calculus to evaluate the PDF of Y_n:

$$f_{Y_n}(y) = \frac{d}{dy}\left[\int_{-\infty}^{\infty} [F_X(y)]^n \, dy\right] = n[F_X(y)]^{n-1} f_X(y)$$

Now, let's go about doing something similar to find Y_1, the *minimum* order statistic. Similar to Y_n, the minimum order statistic is (relatively) easy to compute because it will be related to all of the other random variables in the set in the same way. That is, it will be less than all of the other random variables:

$$Y_1 = X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$$

The PDF of this order statistic is given as:

$$\begin{aligned}
P(Y_1 > y) &= P(\min\{X_1, X_2, \ldots, X_n\} > y) \\
&= P(X_1 > y, X_2 > y, \ldots, X_n > y) \\
&= P(X_1 > y) \cdot P(X_2 > y) \cdots P(X_n > y) \\
&= [1 - P(X_1 \leq y)] \cdot [1 - P(X_2 \leq y)] \cdots [1 - P(X_n \leq y)] \\
&= [1 - F_{X_1}(y)] \cdot [1 - F_{X_2}(y)] \cdots [1 - F_{X_n}(y)] \\
&= [1 - F_X(y)]^n
\end{aligned}$$

Again, the final simplification was made possible because all of $X_1, \ldots, X_n$ are identically distributed. Now, we can compute the CDF of Y_1 as

$$F_{Y_1}(y) = 1 - P(Y_1 > y) = 1 - [1 - F_X(y)]^n$$

and using the fundamental theorem of calculus will afford the PDF:

$$\begin{aligned}
f_{Y_1}(y) &= \frac{d}{dy}\left[\int_{-\infty}^{\infty} 1 - [1 - F_X(y)]^n \, dy\right] = -n[1 - F_X(y)]^{n-1}(-f_X(y)) \\
&= n[1 - F_X]^{n-1} f_X(y)
\end{aligned}$$

Let's consider an example to show how we would apply these new formulas we just derived: Consider the random variables $X_1, \ldots, X_n \sim \text{iidExp}(\lambda)$. We already know the PDF and CDF of these random variables, namely,

$$\begin{aligned}
f_X(x) &= \lambda e^{-\lambda x}, \quad x > 0 \\
F_X(x) &= 1 - e^{-\lambda x}, \quad x > 0
\end{aligned}$$

Finding the minimum and maximum order statistics is as easy as plugging in to the equations we derived above. The PDF of $Y_1 = X_{(1)}$ is given by

$$f_{Y_1}(y) = n[1 - F_X]^{n-1} f_X(y) = n[1 - (1 - e^{-\lambda y})]^{n-1}(\lambda e^{-\lambda y}) = n\lambda e^{-(n\lambda)y}, \quad y > 0$$

Thus, we see that $Y_1 \sim \text{Exp}(n\lambda)$. We'll go about computing the maximum order statistic in the same manner, by plugging and chugging. The PDF of $Y_n = X_{(n)}$ is given by:

$$f_{Y_n}(y) = n[F_X(y)]^{n-1} f_X(y) = n[1 - e^{-\lambda x}]^{n-1}(\lambda e^{-\lambda x}) = n\lambda e^{-\lambda x}[1 - e^{-\lambda x}]^{n-1}$$

Sadly, this expression cannot be easily simplified and so we'll leave this as our final answer.

Now, let's shift our attention to the jth order statistic of an iid set of random variables. What will the PDF of $Y_j = X_{(j)}$ for $1 \leq j \leq n$ be? First, recall the limit definition for the PDF of a continuous random variable:

$$f_{Y_j}(y) = \lim_{h \to 0^-} \frac{P(y - h < Y_j \leq y)}{h} = \lim_{h \to 0^-} \frac{F_{Y_j}(y) - F_{Y_j}(y - h)}{h}$$

Now, let's fix a value for $0 < h < 1$ and look at the following event:

$$\begin{aligned}(y - h < Y_j \leq y) = &\left\{ \text{exactly } j-1 \text{ of the } X_1, \ldots, X_n \text{ are in } (-\infty, y-h] \right\} \\ &\cap \left\{ \text{exactly one of the remaining } X_i\text{'s is in } (y-h, y] \right\} \\ &\cap \left\{ \text{the remaining } X_i\text{'s are in } (y, \infty) \right\}\end{aligned}$$

In other words, we are interested in the event that on some small interval $(y - h, y]$, there is exactly one random variable X_i. In terms of the other order statistics already mentioned, Y_1 and Y_n, Y_1 would be the event that there are no events on the interval $(-\infty, y - h]$ and Y_n would be the event that there are no events on the interval (y, ∞). Let's break this event into pieces, based on the interval in which they take place:

$(-\infty, y - h]$: For these $j - 1$ events, we are interested in the probability that $P(X_i \leq y)$. Notice that this expression is just the CDF of the random variable. Thus, the total probability of this event occurring within this interval is

$$[P(X_i \leq y)]^{j-1} = [F_{X_i}(y)]^{j-1}$$

$(y - h, y]$: For this single event, we are interested in the probability of a single random variable X_i being on the interval. This is given by:

$$P(X_i \leq y) - P(X_i \leq y - h) = F_{X_i}(y) - F_{X_i}(y - h)$$

(y, ∞): For the remaining $n - j$ events, we are interested in the probability of the *complement* of the CDF. This is because we desire that all of the remaining random variables be greater than the X_i we are interested in. Thus, the probability is:

$$[1 - P(X_{i+1} \leq y)]^{n-j} = [1 - F_{X_i+1}(y)]^{n-j}$$

Because all of the random variables are independent we can multiply them together to find the total probability. Before we do that, however, remember to account for all of the possible arrangements of the random variables i.e., given a particular X_i we need to count the number of ways to arrange all of the other variables less than and greater than it. To do this we can use the multinomial coefficient. Now, plugging into our limit definition for the PDF:

$$\begin{aligned}&= \lim_{h \to 0^-} \frac{\binom{n}{j-1,1,n-j}[F_{X_i}(y)]^{j-1} \cdot [F_{X_i}(y) - F_{X_i}(y-h)] \cdot [1 - F_{X_i+1}(y)]^{n-j}}{h} \\ &= \binom{n}{j-1, 1, n-j}[F_{X_i}(y)]^{j-1} \cdot [f_{X_i}(y)] \cdot [1 - F_{X_i+1}(y)]^{n-j} \\ \star\ f_{Y_j}(y) &= \binom{n}{j-1, 1, n-j}[F_X(y)]^{j-1}[1 - F_X(y)]^{n-j} f_X(y)\ \star\end{aligned}$$

With that, we've found the PDF for the jth order statistic of a set of independent and identically distributed variables.

As a brief aside, let's consider the joint PDF of the order statistics. For a set of $X_1, \dots, X_n \sim \text{iid} f$ random variables with order statistics $Y_1, \dots, Y_n$, the joint PDF of the order statistics is given by:

$$f_{Y_1,\dots,Y_n}(y_1, \dots, y_n) = n! f_{Y_1}(y_1) f_{Y_2}(y_2) \cdots f_{Y_n}(y_n)$$

We can find the marginal distribution of Y_i for $i = 1, 2, \dots, n$ by integrating out the contributions from the other order statistics. Below are some examples:

$$f_{Y_2}(y_2) = \int n! f_{Y_1} f_{Y_2} \cdots f_{Y_n} \, dy_1 dy_3 \dots dy_n$$

$$f_{Y_3,Y_4}(y_3, y_4) = \int n! f_{Y_1} f_{Y_2} \cdots f_{Y_n} \, dy_1 dy_2 dy_5 \dots dy_n$$

10 Cheat Sheet

10.1 Probability Functions

Distribution	Functional Form
$X \sim \mathrm{d}\mathcal{U}(a,b)$	$P(X=x)=\frac{1}{n} \quad \text{for } x=x_1,\dots,x_n$
$X \sim \text{Bernoulli}(p)$	$P(X=x)=p^x(1-p)^x \quad \text{for } x=0,1$
$X \sim \text{Binomial}(n,p)$	$P(X=x)=\binom{n}{x}p^x(1-p)^{n-x} \quad \text{for } x=0,1,2,\dots,n$
$X \sim \text{NB}(r,p)$	$P(X=x)=\binom{x-1}{r-1}p^r(1-p)^{x-r} \quad \text{for } x=r,r+1,r+2,\dots$
$X \sim \text{Multinomial}(n,r)$	$P(X_1=x_1,\dots,X_r=x_r)=\binom{n}{x_1,\dots,x_r}p_1^{x_1}\cdots p_r^{x_r}$ for $x=0,1,2,\dots$
$X \sim \text{Geometric}(p)$	$P(X=x)=p(1-p)^{x-1} \quad \text{for } x=1,2,3,\dots$
$X \sim \text{Hypergeometric}(n,M,N)$	$P(X=x)=\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \quad \text{for } x=0,1,2,\dots,M$
$X \sim \text{Poisson}(\lambda)$	$P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!} \quad \text{for } x=0,1,2,\dots$
$X \sim \text{Logarithmic}(p)$	$P(X=x)=-\frac{p^x}{x\ln(1-p)} \quad \text{for } x=1,2,3,\dots$
$X \sim \mathcal{U}(a,b)$	$f_X(x)=\frac{1}{b-a} \quad \text{for } x\in[a,b]$
$X \sim \text{Exp}(\lambda)$	$f_X(x)=\lambda e^{-\lambda x} \quad \text{for } x\geq 0$
$X \sim \text{Gamma}(\alpha,\beta)$	$f_X(x)=\frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)} \quad \text{for } x>0$
$X \sim \mathcal{N}(\mu,\sigma^2)$	$f_X(x)=\frac{e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}}{\sigma\sqrt{2\pi}} \quad \text{for } -\infty<x<\infty$
$X \sim \text{Log-}\mathcal{N}(\mu,\sigma^2)$	$f_X(x)=\frac{e^{-\frac{1}{2}(\frac{\ln x-\mu}{\sigma})^2}}{x\sigma\sqrt{2\pi}} \quad \text{for } 0<x<\infty$
$X \sim \chi^2_n$	$f_X(x)=\frac{x^{(n/2)-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)} \quad \text{for } x>0$
$X \sim \text{Beta}(\alpha,\beta)$	$f_X(x)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} \quad \text{for } 0\leq x\leq 1$
$X \sim \text{F}_{n,m}$	$f_X(x)=\frac{\Gamma[(n+m)/2]}{\Gamma(n/2)\Gamma(m/2)}(\frac{n}{m})^{n/2}x^{(n/2)-1}(1+\frac{nx}{m})^{-(n+m)/2} \quad \text{for } x>0$
$X \sim \text{Student's t}_m$	$f_X(x)=\frac{\Gamma[(m+1)/2]}{\sqrt{m\pi}\Gamma(m/2)}(1+\frac{x^2}{m})^{-(m+1)/2} \quad \text{for } x>0$
$X \sim \mathcal{L}(\lambda)$	$f_X(x)=\frac{\lambda}{2}e^{-\lambda\lvert x-c\rvert} \quad \text{for } c\in\mathbb{R},\lambda>0,-\infty<x<\infty$
$X \sim \mathcal{BVN}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)$	$f(x_1,x_2)=\frac{\exp\left\{\frac{-1}{2(1-\rho^2)}\left[\left(\frac{X_1-\mu_1}{\sigma_1}\right)^2-2\rho\left(\frac{X_1-\mu_1}{\sigma_1}\right)\left(\frac{X_2-\mu_2}{\sigma_2}\right)+\left(\frac{X_2-\mu_2}{\sigma_2}\right)^2\right]\right\}}{\sigma_1\sigma_2 2\pi\sqrt{1-\rho^2}}$ $(x,y)\in\mathbb{R}^2$
$X \sim \text{Pareto}(\alpha,\lambda)$	$f_X(x)=\frac{\alpha\lambda^\alpha}{x^{\alpha+1}} \quad \text{for } x>\lambda$

10.2 Expected Values, Variances, and Moment Generating Functions

Distribution	$E(X)$	$\text{Var}(X)$	$M(\theta)$
$X \sim \text{d}\mathcal{U}(a, b)$	$\frac{1}{n}\sum_{i=1}^{n} x_i$	$\frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2\right]$	$\frac{1}{n}\sum_{i=1}^{n} e^{\theta x_i}$
$X \sim \text{Bernoulli}(p)$	p	$p(1-p)$	$1-p+pe^{\theta}, \theta \in \mathbb{R}$
$X \sim \text{Binomial}(n, p)$	np	$np(1-p)$	$(1-p+pe^{\theta})^n$
$X \sim \text{NB}(r, p)$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$	$\left(\frac{pe^{\theta}}{1-(1-p)e^{\theta}}\right)^r$ $\theta < -\ln(1-p)$
$X \sim \text{Multinomial}(n, r)$	$E(X_i) = np_i$	$\sigma_i^2 = np_i(1-p_i)$ $\sigma_{i,j} = -np_ip_j, i \neq j$	
$X \sim \text{Geometric}(p)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^{\theta}}{1-(1-p)e^{\theta}}$ $\theta < -\ln(1-p)$
$X \sim \text{Hypergeometric}(n, M, N)$	$\frac{nM}{N}$	$\frac{nM}{N}\left(1-\frac{M}{N}\right)\left(\frac{N-1}{N-n}\right)$	
$X \sim \text{Poisson}(\lambda)$	λ	λ	$e^{\lambda(e^{\theta}-1)}$
$X \sim \text{Logarithmic}(p)$	$-\frac{p}{(1-p)\ln(1-p)}$	$-\frac{p^2+p\ln p}{(1-p)^2[\ln(1-p)]^2}$	$\frac{\ln(1-pe^{\theta})}{\ln(1-p)}$ $\theta < -\ln p$

Distribution	$E(X)$	$\mathrm{Var}(X)$	$M(\theta)$
$X \sim \mathcal{U}(a,b)$	$\frac{1}{2}(b+a)$	$\frac{1}{12}(b-a)^2$	$\frac{1}{\theta(b-a)}(e^{\theta b}-e^{\theta a})$
$X \sim \mathrm{Exp}(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$(1-\frac{\theta}{\lambda})^{-1}$ $\theta<\lambda$
$X \sim \mathrm{Gamma}(\alpha,\beta)$	$\alpha\beta$	$\alpha\beta^2$	$(1-\beta\theta)^{-\alpha}$ $\theta<1/\beta$
$X \sim \mathcal{N}(\mu,\sigma^2)$	μ	σ^2	$\exp\{\mu\theta+\frac{\sigma^2\theta^2}{2}\}$
$X \sim \text{Log-}\mathcal{N}(\mu,\sigma^2)$	$e^{\mu+\frac{\sigma^2}{2}}$	$[e^{\sigma^2}-1]e^{2\mu+\sigma^2}$	
$X \sim \chi^2_n$	n	$2n$	$(1-2\theta)^{-\frac{n}{2}}$ $\theta<\frac{1}{2}$
$X \sim \mathrm{Beta}(\alpha,\beta)$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	
$X \sim \mathrm{F}_{n,m}$	$\frac{m}{m-2}$ $m>2$	$\frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$ $m>4$	
$X \sim \text{Student's } \mathrm{t}_m$	0 $m>1$	$\frac{m}{m-2}$ $m>2$	
$X \sim \mathcal{L}(\lambda)$	c	$\frac{2}{\lambda^2}$	$\frac{e^{c\theta}}{1-(\theta/\lambda)^2}$ $-\lambda<\theta<\lambda$
$X \sim \mathcal{BVN}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)$	$E(X_1)=\mu_1$ $E(X_2)=\mu_2$	$\mathrm{Var}(X_1)=\sigma_1^2$ $\mathrm{Var}(X_2)=\sigma_2^2$ $\mathrm{corr}(X_1,X_2)=\rho$	
$X \sim \mathrm{Pareto}(\alpha,\lambda)$	$\frac{\alpha\lambda}{\alpha-1}$ $\alpha>1$	$\frac{\alpha\lambda^2}{(\alpha-1)^2(\alpha-2)}$ $\alpha>2$	

11 Afterword