Some techniques in mathematical statistics- generating functions

Peter Haggstrom
mathsatbondibeach@gmail.com
https://gotohaggstrom.com

September 30, 2023

1 Background

Mathematical statistics is a highly technical mathematical discipline. It involves fundamental probability theory encompassing the standard frequentist approach as well as the Bayesian approach. There is a huge superstructure of analysis and calculus which underpins mathematical statistics and there are many other areas of mathematics which also have an impact. Then there is the whole area of descriptive statistics with dimensions such as sample design, inference and so on. The Russians have been a driving influence in probability theory from the days of Chebychev and Bernstein through to names like Smirnov, Glivenko, Lyapunov, Markov and, of course, that giant Kolmogorov. In more recent times Vladimir Vapnik has had an important role in the development of applying statistical methods to machine learning. It was Glivenko and Cantelli who proved that the empirical distribution function always converges to the actual distribution function and Kolmogorov found the asymptotically exact rate of this convergence, the rate of which turns out to be exponentially fast and independent of the unknown distribution function. That in itself suggests the technical depth of the discipline. To be truly proficient in the discipline you need to know classical probability theory (including combinatorics, parametric and non-parametric statistical theory, real analysis, complex analysis, functional analysis including harmonic analysis, Lebesgue integration theory, Fourier theory, Laplace transform theory, asymptotic theory, finite difference theory, ordinary and partial differential equations, stochastic calculus, matrix theory etc. Very few people actually embrace all these skill areas. The need for rigour in this discipline is underscored by the fact that the American Statistical Association has had to issue guidance on the interpretation of $p-$values given the scale of the bogus interpretations that abound in medicine, psychology and so on: https://amstat.tandfonline.com/doi/epdf/10.1080/00031305.2016.1154108?needAccess=true&role=button

Just for the record here is what the Glivenko-Cantelli theorem says. Let's suppose that X_i for $i = 1, 2, \ldots, n$ are identical independently distributed random variable with distribution function F on $\mathbb{R}$. The empirical distribution function is the function of F defined by:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{1 \leq i \leq n} I\{X_i \leq x\} \tag{1}$$

Where I is the indicator set function. The theorem then says that:

$$\sup_{x\in\mathbb{R}}|\hat{F}_n(x) - F(x)| \to 0 \quad \text{almost surely} \tag{2}$$

So the empirical distribution is a reasonable esimate of the abstract distribution function. A pretty fundamental result.

The purpose of this article is to deal with a couple of mathematical statistics problems and highlight the use of generating functions, of both the moment and factorial species. It is by no means exhaustive! Here are two "simple" problems which reflect the depth of the subject.

Problem 1

In a TedX lecture Australian-British mathematical statistician Sir Peter Donnelly, FRS, who works at Oxford University and is the CEO of Genomics PLC, asked the audience to vote on the following proposition. Suppose we toss an unbiased coin and we are looking for the expected number of tosses to get "HTH" versus the expected number of tosses to get "HTT". Is the expected number of tosses of "HTH" greater than, less than or equal to the expected number of tosses to get "HTT" ? Here is the link to the lecture `https://www.youtube.com/watch?v=kLmzxmRcUTo&t=361s`

The audience said the expected numbers were the same when in fact the expected number of tosses for "HTH" is 10 and for "HTT" it is 8. Even highly qualified mathematicians have got this wrong and it certainly is not intutitive even though the probabilities of both runs is $\frac{1}{8}$. To prove Donnelly's assertion requires some sophisticated analysis which is set out in the attached paper which also deals with how poorly the legal system deals with statistical/probabilistic arguments: `https://www.gotohaggstrom.com/Fooling%20juries%20with%20statistics.pdf`

The late Peter Gavin Hall, FRS, was another notable and highly cited Australian mathematical statistician who made significant contributions to statistical theory.

Problem 2

In his famous textbook [3] William Feller describes at pages 86-88 a computer simulation of 10,000 coin tosses. He notes that "most people feel surprised by the length of the intervals between successive crossings of the axis. As a matter of fact, the graph represents a rather mild case history and was chosen as the mildest among the available records. A more startling example is obtained by looking at the same graph in the *reverse* direction; that is , reversing the order in whihc the 10,000 trials actually occurred. Theoretically, the series as graphed and the reversed series are equally legitimate as representative of an ideal random walk." He then produces a table of the reversed graph which shows that starting from the origin the path stays on the negative side for 9930 steps and on the positive side for 70 steps. He says: "This looks absurd, and yet the probability that in 10,000 tosses of a perfect coin the lead is at one side for more than 9930 trials and at the other for fewer than 70 exceeds $\frac{1}{10}$. In other words, on the average *one record out of ten will look worse than the one just described.* By constrast, the probability of a balance better than in the graph is only 0.072. Feller also noted that "sampling of expert opinion revealed that even trained statisticians expect much more than 78 changes of sign in 10,000 trials, and nobody counted on the possibility of only 8 changes of sign. Actually the probability of not more than 8 changes of sign exceeds 0.14 , whereas the probability of of

more than 78 changes of sign is about 0.12. As far as the number of changes of sign is concerned the two records stand on a par and theoretically, neither should cause surprise. If they seem startling, this is due to our faulty intuition and to our having been exposed to too many vague references to a mysterious "law of averages".

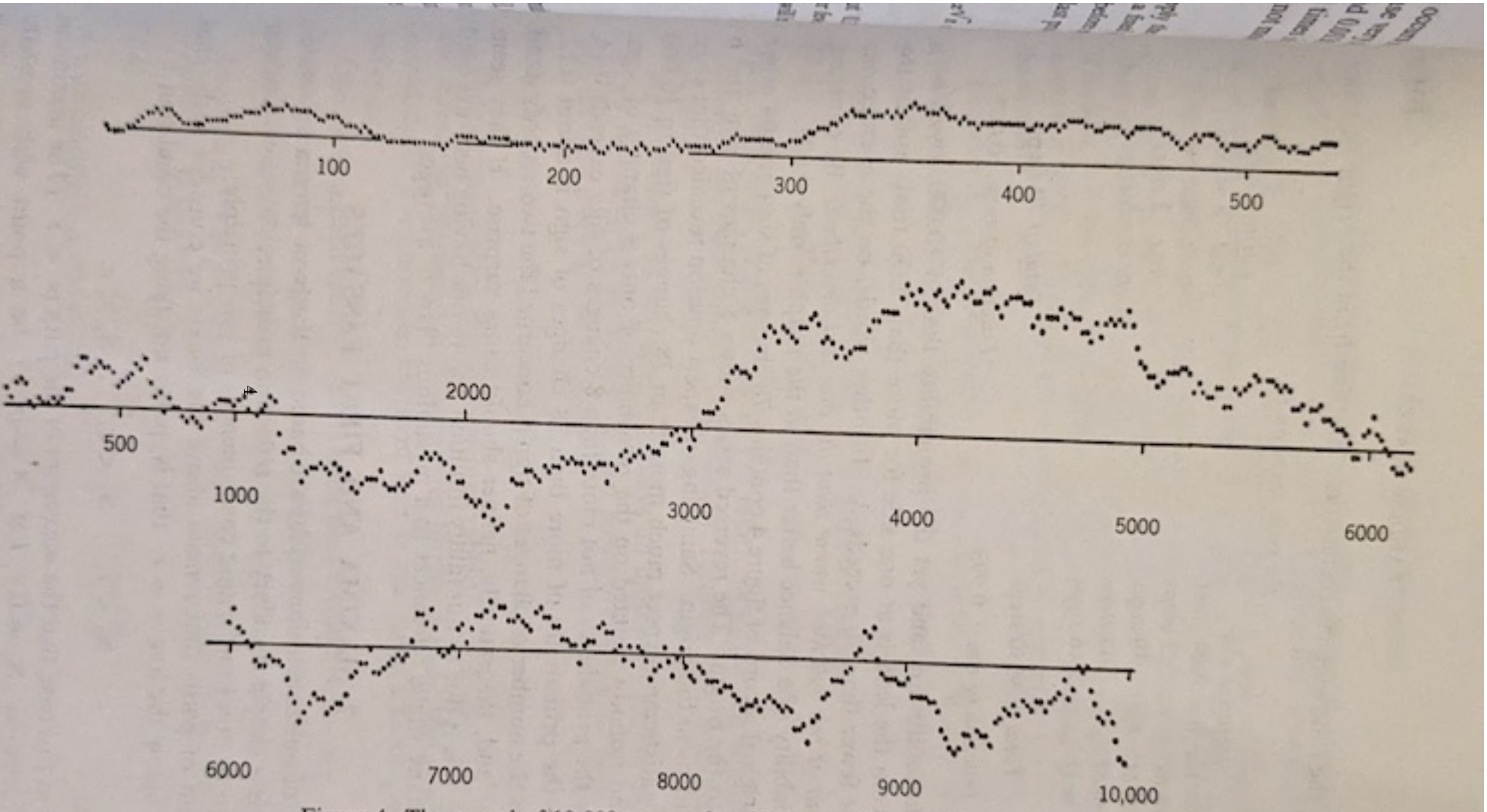


Figure 4. The record of 10,000 tosses of an ideal coin (described in section 6).

The arc sine law which underpins the behaviour described by Feller is explained in more detail in this paper: https://www.gotohaggstrom.com/What%20do%20schmucks%20and%20the%20arc%20sine%20law%20have%20in%20common.pdf

2 Summary

Suppose X is a random variable with a given distribution, and that, for some real number $h > 0$, $\mathbb{E}[e^{sX}]$ exists for every $s \in (-h, h)$, then $M(s)$ is called the moment generating function (mgf) of X:

$$\boxed{M(s) = \mathbb{E}[e^{sX}]}$$

When X is discrete the mgf is:

$$\boxed{M(s) = \sum_{x_i} e^{sx_i}\mathbb{P}[X = x_i]}$$

When X is continuous the mgf is:

$$\boxed{M(s) = \int_{-\infty}^{\infty} e^{sx} f(x)\, dx}$$

Example

For a Bernoulli distribution $M(s) = e^{s.0}\mathbb{P}[X = 0] + e^{s.1}\mathbb{P}[X = 1] = (1 - p) + pe^s$ which holds for $-\infty < s < \infty$.

How moments are generated

$$\boxed{M(s) = \sum_{r=0}^{\infty} \mathbb{E}[X^r]\frac{s^r}{r!}}$$

$\boxed{\mathbb{E}[X^r] = M^{(r)}(0)}$ for $r = 1, 2, \ldots$

Transformation results

Suppose X has mgf $M_X(s)$ and let $Y = aX + b$ for real a, b. Then the mgf of Y is:

$$\boxed{M_Y(s) = M_{aX+b}(s) = e^{bs}M_X(as)}$$

If X and Y are independent with mgfs $M_X(s)$ and $M_Y(s)$ then for any constants a, b:

$$\boxed{M_{aX+bY}(s) = M_X(as).M_Y(bs)}$$

Reproductive properties

Suppose that $X_1, X_2, \ldots, X_n$ is any finite collection of independent random variables belonging to a family of distributions, then the family is said to have the reproductive property if the distribution $X_1 + X_2 + \ldots X_n$ also belongs to the family.

The "usual suspects" with this property are the binomial distributions with parameter p, the Poisson, the normal and chi-square distributions.

For example, if $X_1, X_2, \ldots X_r$ are independent normal variables such that X_i is $N(\mu_i, \sigma_i^2)$ for $i = 1, 2, \ldots r$ then $X_1 + X_2 + \cdots + X_r$ is $N(\sum_{i=1}^r \mu_i, \sum_{i=1}^r \sigma_i^2)$.

Factorial moment generating functions or probability generating functions

The factorial moment generating function (fmgf) is defined as:

$$\boxed{g(s) = \mathbb{E}[s^X]}$$

The derivatives of $g(s)$ are given by:

$$\boxed{g^{(r)}(s) = \mathbb{E}[X(X-1)\ldots(X-r+1)s^{X-r}]}$$

and so

$$\boxed{g^{(r)}(1) = \mathbb{E}[X(X-1)\ldots(X-r+1)]}$$

For non-negative integral values of the variable we have that the probability that $X = r$ is:

$$\boxed{\mathbb{P}[X = r] = \frac{g^{(r)}(0)}{r!} \text{ for } r = 0, 1, 2\ldots}$$

Relationship between the mgf and fmgf

$$\boxed{g(s) = M(\ln s)}$$

$$\boxed{M(s) = g(e^s)}$$

Uniqueness theorem

If two random variables have the same mgf's then they have the same distribution and conversely.

For instance, if a random variable X has the mgf $M(s) = e^{2(e^s-1)}$ then we can conclude that X has the Poisson distribution with parameter $\lambda = 2$.

3 Background on generating functions

The use of generating functions is fundamental to probability theory. One of the deepest and most useful properties of a moment generating function is that when it exists it is unique. Thus if two

random variables have the same moment generating functions then they have the same distribution, and conversely. In other words, if X and Y are two random variables, then $M_X(s) = M_Y(s)$ for $s \in (-h, h)$ ($h > 0$) if and only if $F_X(u) = F_Y(u)$ for all real u. For example, if by some means we work out that X has the moment generating function $M(s) = e^{2(e^s-1)}$ then we can conclude that X has the Poisson distribution with parameter $\lambda = 2$ and no other distribution can have such a moment generating functions.

It is by no means straight forward to prove this rather remarkable concept. To do it properly you basically have to look to the complex valued analogue of the moment generating function, namely, the characteristic function of a random variable X defined as:

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{itx} f_X(x)\, dx \tag{3}$$

One will see other definitions such as $\phi(t) = \int_{-\infty}^{\infty} e^{itx} \mu(dx)$ where μ is a probability measure or $\mathbb{E}[e^{itX}]$ and so on. Thus the characteristic function is a moment generating function with a complex argument it. In non-probabilistic contexts the characteristic function is called a Fourier transform and there is a vast theoretical edifice that can be called upon to justify inversion of the transform as well as many other properties. This relationship with Fourier theory gives rise to 3 really fundamental properties:

(1) If two variables X, Y (or probability measures) have characteristic fuinctions $\phi_X(t)$ and $\phi_Y(t)$ the the convolution $X * Y$ has characteristic function $\phi_X(t)\, \phi_Y(t)$. This is essentially the probabilistic analogue of the Fourier transform of a convolution being the product of the Fourier transforms.

(2) The characteristic function uniquely determines the distribution so that no information is lost in studying products of characteristic functions for instance.

(3) The pointwise convergence of characteristic functions implies the weak convergence of the corresponding distributions so in a study of asymptotic distributions of sums of independent random variables you can focus on the behaviour of the characteristic functions. This is referred to as a continuity theorem in the sense that a necessary and sufficient condition for $X_n \Rightarrow X$ (weak convergence) is that $\phi_{X_n}(t) \to \phi_X(t)$. For the purposes of this paper I am not going to drill down into the various species of convergence but what I will do is set out some of the proof of the uniqueness property in the Appendix.

It is important to note that there are two main species of generating functions:

(a) Moment generating functions; and

(b) Probability (for integral values of the variable) or factorial moment generating functions.

There is a relationship between the two which will be spelled out below with examples.

The usual way of introducing moment generating functions is to assume that $\mathbb{E}[e^{sX}]$ exists for every value $s \in (-h, h)$ for some $h > 0$. The moment generating function ("mgf") is then defined as:

$$M(s) = \mathbb{E}[e^{sX}] \tag{4}$$

For clarity one can write $M_X(s)$ to emphasis that the mgf is based on the distribution of X.

If the distribution is continuous we write:

$$M(s) = \int_{-\infty}^{\infty} e^{sx} f(x)\, dx \tag{5}$$

where $f(x)$ is the density of the distribution X.

If X is discrete we write:

$$M(s) = \sum_{x_i} e^{sx_i} \mathbb{P}[X = x_i] \tag{6}$$

Standard results using this approach are as follows.

Bernoulli with parameter p

$M(s) = e^{s.0}\mathbb{P}[X = 0] + e^{s.1}\mathbb{P}[X = 1] = (1-p) + pe^s$. Hence the mgf exists for all $-\infty < s < \infty$ and:

$$M(s) = pe^s + 1 - p = pe^s + q \tag{7}$$

where $q = 1 - p$.

$$\boxed{M(s) = pe^s + (1-p) \text{ for } -\infty < s < \infty.} \tag{8}$$

Binomial, B(n,p)

$$\begin{aligned} M(s) =& \sum_{k=0}^{n} e^{sk} \binom{n}{k} p^k (1-p)^{n-k} \\ =& \sum_{k=0}^{n} \binom{n}{k} (pe^s)^k (1-p)^{n-k} \\ =& [pe^s + (1-p)]^n \end{aligned} \tag{9}$$

which holds for all $-\infty < s < \infty$

$$\boxed{M(s) = [pe^s + (1-p)]^n \text{ for } -\infty < s < \infty.} \tag{10}$$

Geometric

$$\begin{aligned} M(s) =& \sum_{k=1}^{\infty} e^{sk}\mathbb{P}[X=k] \\ =& \sum_{k=1}^{\infty} e^{sk}p(1-p)^{k-1} \\ =& pe^s \sum_{k=1}^{\infty}[(1-p)e^s]^{k-1} \\ =& pe^s \sum_{r=0}^{\infty}[(1-p)e^s]^{r} \end{aligned} \tag{11}$$

But $\sum_{r=0}^{\infty}[(1-p)e^s]^r$ is a geometric series which converges to $\frac{1}{1-(1-p)e^s}$ if $(1-p)e^s < 1$.

$$\boxed{M(s) = \frac{pe^s}{1-(1-p)e^s} \text{ if } (1-p)e^s < 1.} \tag{12}$$

Note that for the geometric distribution we are essentially looking at the waiting time for a success. Thus for independent trials where X represents the the number of trials for the first success to occur, the possible values of X are $1, 2, \ldots$ and $\mathbb{P}(X=r) = (1-p)^{r-1}p$ where $0 < p < 1$ is the fixed probability of success. There must be at least r trials because success has to occur r times hence $X = r, r+1, \ldots$.

Negative Binomial Distribution with parameters r, p

Recall that the negative binomial or Pascal distribution is the distribution of independent Bernoulli trials with a constant probability of success p where $0 < p < 1$ and the random variable X is defined to be the number of trials needed for r successes to occur where $r = 1, 2, \ldots$. $X = r + k$ will occur if and only if the r^{th} success occurs on the $(r+k)^{th}$ trial. This means that success occurs $(r-1)$ times in the first $r+k-1$ trials AND success occurs on the $(r+k)^{th}$ trial. Using independence we then have that:

$$\begin{aligned} \mathbb{P}(X = r+k) =& \mathbb{P}[\text{success occurs (r-1) times in the first r+k-1 trials }] \cdot \mathbb{P}[\text{success occurs on the } (r+k)^{th} \text{ trial}] \\ =& \binom{r+k-1}{r-1} p^{r-1}(1-p)^k p \\ =& \binom{r+k-1}{r-1} p^{r}(1-p)^k \end{aligned} \tag{13}$$

for $k = 0, 1, 2, \ldots$

Note that we can manipulate the binomial term as follows to see where the "negative" comes from:

$$\begin{aligned}\binom{r+k-1}{k} =& \frac{(r+k-1)(r+k-2)\dots(r+k-1-(k-1))}{k!} \\ =& \frac{(r+k-1)(r+k-2)\dots(r+1)r}{k!} \\ =& \frac{(-r)(-r-1)\dots(-r-k+1)}{k!}(-1)^k \\ =& (-1)^k\binom{-r}{k}\end{aligned} \tag{14}$$

Thus (13) becomes:

$$\mathbb{P}(X = r+k) = \binom{-r}{k}(-1)^k p^r (1-p)^k \tag{15}$$

for $k = 0, 1, 2, \dots$. Note that as r goes from 1 etc k goes from 0 and so on.

The moment generating function is:

$$\begin{aligned}M(s) =& \sum_{k=0}^{\infty} e^{s(r+k)}\mathbb{P}[x = r+k] \\ =& \sum_{k=0}^{\infty} e^{s(r+k)}\binom{r+k-1}{r-1}p^r\,(1-p)^k \\ =& (pe^s)^r \sum_{k=0}^{\infty}\binom{r+k-1}{r-1}[(1-p)e^s]^k \\ =& (pe^s)^r[1-(1-p)e^s]^{-r}\end{aligned} \tag{16}$$

The last line is justifed on the basis that if $|x| < 1$ then:

$$\sum_{k=0}^{\infty}\binom{r+k-1}{r-1}x^k = (1-x)^{-r} \tag{17}$$

so that the mgf in (16) is valid if $(1-p)e^s < 1$. More detail on how to derive this result is given here: `https://www.gotohaggstrom.com/The%20binomial%20series%20for%20negative%20integral%20exponents.pdf`

$$\boxed{M(s) = \left[\frac{pe^s}{1-(1-p)e^s}\right]^r \text{ if } (1-p)e^s < 1.} \tag{18}$$

Poisson distribution with parameter λ

Recall that the Poisson distribution with parameter λ is defined as follows:

$$\mathbb{P}(X = k) = \frac{e^{-\lambda}\,\lambda^k}{k!} \tag{19}$$

for $k = 0, 1, 2, \ldots$. Note that the parameter can be more generally written as λt for a time interval of length t . On that basis in (19) $t = 1$

The mgf is as follows:

$$\begin{aligned} M(s) =& \sum_{k=0}^{\infty} e^{sk}\mathbb{P}[X = k] \\ =& \sum_{k=0}^{\infty} e^{sk}\frac{e^{-\lambda}\,\lambda^k}{k!} \\ =& e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^s)^k}{k!} \\ =& e^{-\lambda}e^{\lambda e^s} \\ =& e^{\lambda(e^s-1)} \end{aligned} \tag{20}$$

for all $-\infty < s < \infty$.

$$\boxed{M(s) = e^{\lambda(e^s-1)} \text{ for all } -\infty < s < \infty.} \tag{21}$$

Uniform distribution on $[a, b]$

This is of course a continuous distribution (actually absolutely continuous).The probability density function is $f(x) = \frac{1}{b-a}$ for $x \in [a, b]$ and 0

$$M(s) = \int_a^b e^{sx}\,f(x)\,dx = \int_a^b e^{sx}\,\frac{1}{b-a}\,dx = \frac{e^{bs} - e^{as}}{s(b-a)} \tag{22}$$

if $s \neq 0$. Note that $M(0) = 1$ Hence:

$$\boxed{M(s) = \begin{cases} \frac{e^{bs}-e^{as}}{s(b-a)} & s \neq 0 \\ 1 & s = 0 \end{cases}} \tag{23}$$

Standard normal distribution

Recall that the probability density functiion is $f(x) = \frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}$ and so the mgf is:

$$\begin{aligned} M(s) =& \int_{-\infty}^{\infty} e^{sx} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\, dx \\ =& \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-(x^2-2sx)}{2}}\, dx \\ =& \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{[\frac{-(x^2-2sx+s^2)}{2}]+\frac{s^2}{2}}\, dx \\ =& e^{\frac{s^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-(x-s)^2}{2}}\, dx \\ =& e^{\frac{s^2}{2}} \end{aligned} \tag{24}$$

since $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-(x-s)^2}{2}}\, dx$ represents the area under the normal distribution $N(s,1)$ and so must equal 1.

$$\boxed{M(s) = e^{\frac{s^2}{2}},\ -\infty < s < \infty} \tag{25}$$

To work out the mgf for $N(\mu, \sigma^2)$ one can evaluate the integral $\int_{-\infty}^{\infty} e^{sx} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}\, dx$ or employ a transformation approach. Thus suppose X has mgf M_X and let $Y = aX + b$ where a, b are any real numbers. Then we have:

$$\begin{aligned} M_{aX+b}(s) =& \mathbb{E}[e^{(aX+b)s}] \\ =& \mathbb{E}[e^{bs}.e^{asX}] \\ =& e^{bs}\mathbb{E}[e^{asX}] \\ =& e^{bs} M_X(as) \end{aligned} \tag{26}$$

Note that e^{bs} is just a constant and can be taken outside the expectation operator.

Using this principle we can work out the mgf of $N(\mu, \sigma^2)$ by noting that if $X = \frac{Y-\mu}{\sigma}$ then X is $N(0,1)$. This can be seen as follows. The distribution function for X is $F_X(t) = \mathbb{P}[X \le t] = \mathbb{P}[\frac{Y-\mu}{\sigma} \le t] = \mathbb{P}[Y \le \sigma t + \mu]$. But because Y is $N(\mu, \sigma^2)$ we then have:

$$\begin{aligned} \mathbb{P}[Y \le \sigma t + \mu] &= \int_{-\infty}^{\sigma t+\mu} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(y-\mu)^2}{2\sigma^2}}\, dy \\ &= \int_{-\infty}^{t} \frac{\sigma}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2}}\, dx \text{ using the substitution } x = \frac{y-\mu}{\sigma} \\ &= \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\, dx \end{aligned} \tag{27}$$

and this demonstrates that X is $N(0,1)$. Because $M_X(s) = e^{\frac{s^2}{2}}$ we have$M_{\sigma X}(s) = M_X(\sigma s) = e^{\frac{(\sigma s)^2}{2}}$. Finally we have that:

$$M_Y(s) = M_{\sigma X+\mu}(s) = e^{\mu s} e^{\frac{\sigma^2 s^2}{2}} = e^{\mu s + \frac{\sigma^2 s^2}{2}} \tag{28}$$

So for Y $N(\mu, \sigma^2)$:

$$\boxed{M_Y(s) = e^{\mu s + \frac{\sigma^2 s^2}{2}} \ , -\infty < s < \infty} \tag{29}$$

If X and Y are independent with mgfs $M_X(s)$ and $M_Y(s)$ then for any constants a, b we have:

$$\boxed{M_{aX+bY}(s) = M_X(as).M_Y(bs)} \tag{30}$$

This is proved by using the fact that $\mathbb{E}$ preserves independence ie if X and Y are independent then so are $\mathbb{E}[e^{saX}]$ and $\mathbb{E}[e^{sbY}]$. Clearly the logic can be replicated so that the mgf of a finite number of independent random variables is equal to the product of their mgfs.

Gamma distribution with parameters $\lambda > 0$ and $p > 0$

Revcall that the gamma distribution is characterised by the following probability density function:

$$f(x) = \begin{cases} \frac{\lambda^p}{\Gamma(p)} x^{p-1} e^{-\lambda x} & , x > 0 \\ 0 & \text{elsewhere} \end{cases} \tag{31}$$

Hence the mgf is as follows assuming $\lambda > s$ and the substitution $(\lambda - s)x = y$ is used:

$$\begin{aligned} M(s) =& \int_0^\infty e^{sx} \frac{\lambda^p}{\Gamma(p)} x^{p-1} e^{-\lambda x} \, dx \\ =& \frac{\lambda^p}{\Gamma(p)} \int_0^\infty x^{p-1} e^{-(\lambda - s)x} \, dx \\ =& \frac{\lambda^p}{\Gamma(p)} \int_0^\infty \left(\frac{y}{\lambda - s} \right)^{p-1} e^{-y} \frac{dy}{\lambda - s} \\ =& \left(\frac{\lambda}{\lambda - s} \right)^p \frac{1}{\Gamma(p)} \int_0^\infty y^{p-1} e^{-y} \, dy \\ =& \left(\frac{\lambda}{\lambda - s} \right)^p \end{aligned} \tag{32}$$

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

Thus if X has a gamma distribution with parameters $\lambda > 0$ and $p > 0$ then:

$$\boxed{M(s) = \left(\frac{\lambda}{\lambda - s}\right)^p \text{ if } s < \lambda}$$

Exponential distribution with parameter λ

The exponential distribution is a special case of the gamma distribution with $p = 1$. It has pdf:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

When we put $p = 1$ in (32) we get its mgf as :

$$\boxed{M(s) = \frac{\lambda}{\lambda - s} \text{ if } s < \lambda} \tag{33}$$

4 Generating functions for integral valued random variables

It is useful to give some background about generating functions which go as far back as De Moivre and Laplace. The following is based on Feller ([3] Chapter XI). Suppose we have a sequence of real numbers $a_0, a_1, a_2, \ldots$ which could represent a sequence of some random experiment such as coin tossing, in which case the sequence could be something like $0, 0, 1, 1, 0, 1, \ldots$ for instance. We construct a sum:

$$A(s) = a_0 + a_1 s + a_2 s^2 + \ldots \tag{34}$$

The variable s is just a placeholder and has no intrinsic significance, since the focus is on the coefficients. If we assume that the infinite series $A(s)$ converges in some interval $-s_0 < s < s_0$ then $A(s)$ is called the generating function of the sequence $\{a_j\}$. If the sequence $\{a_j\}$ is bounded by some number $B > 0$ then $A(s)$ is dominated by $B(1 + s + s^2 + \ldots)$ which is just a geometric series and so it will converge for $|s| < 1$. Thus $A(s)$ will converge by comparison principles. If the $a_j = 1$ for all j then $A(s) = \frac{1}{1-s}$. If we had a sequence of the following form $(0, 0, 1, 1, 1, 1, \ldots)$ we can see that is $A(s) = 0.1 + 0.s + 1.s^2 + 1.s^3 + \cdots = s^2 + s^3 + s^4 + \ldots$ so that $A(s) = \frac{s^2}{1-s}$. The generating sequence for $a_j = \frac{1}{j!}$ is simply e^s. Similarly, for fixed n the sequence $a_j = \binom{n}{j}$ has generating function $(1+s)^n$. If we are interested in the number scored in a throw of a perfect die, the probability distribution of that random variable has a generating function $\frac{s+s^2+s^3+s^4+s^5+s^6}{6}$.

If $\boldsymbol{X}$ is a random variable assuming values $0, 1, 2, \ldots$ and we let $\mathbb{P}\{\boldsymbol{X} = j\} = p_j$, and $\mathbb{P}\{\boldsymbol{X} > j\} = q_j$ then it follows that:

$$q_k = p_{k+1} + p_{k+2} + \ldots \quad \text{for } k \geq 0 \tag{35}$$

The generating functions for the sequences $\{p_j\}$ and $\{q_j\}$. are:

$$\begin{aligned} P(s) =& p_0 + p_1 s + p_2 s^2 + p_3 s^3 + \dots \\ Q(s) =& q_0 + q_1 s + q_2 s^2 + q_3 s^3 + \dots \end{aligned} \tag{36}$$

First note that $P(1) = 1$ since the sum of the probabilities must be 1 and thus $P(s)$ converges absolutely certainly for $-1 \leq s \leq 1$. The coefficients of $Q(s)$ are strictly less than 1 and so we can be sure that $Q(s)$ converges (absolutely) at least on the open interval $-1 < s < 1$. Note that from the theory of power series, there is a radius of convergence $R \geq 0$ such that the sum converges absolutely if $|s| < R$ and diverges if $|s| > R$. Moreover the sum is uniformly convergent on sets like $\{|s| \leq R'\}$ for any $R' < R$. The sum can be differentiated and integrated term by term any number of times when $|s| < R'$.

The use of generating functions often involves equating coefficients of the relevant series representations. This is well known in the context of combinatorial problems. For instance the combinatorial formula:

$$\sum_i \binom{n}{i}^2 = \binom{2n}{n} \tag{37}$$

can be obtained directly by equating coefficients of s^0 in the identity:

$$(1+s)^n(1+s^{-1})^n = s^{-n}(1+s)^{2n} \tag{38}$$

It is a basic theorem in generating functions that for $-1 < s < 1$ we have:

$$Q(s) = \frac{1 - P(s)}{1 - s} \tag{39}$$

To prove this note that the coefficient of s^n in $(1-s) \centerdot Q(s) = q_n - q_{n-1} = -p_n$ when $n \geq 1$ (see equation (35)). Hence the coefficient of s^n in $1 - P(s) = -p_n$ as required. Now when $n = 0$ we have from (4) that $q_0 = p_1 + p_2 + \dots = 1 - p_0$. Thus the theorem is proved.

Derivatives of generating functions are also a fundamental building block. Formally differentiating $P(s)$ we have that:

$$P'(s) = \sum_{k=1}^{\infty} k p_k s^{k-1} \tag{40}$$

The theory of real power series ensures that within the interval of convergence(ie on any interval which lies entirely within the interval of convergence) the series is absolutely convergent and uniformly convergent. Moreover such a power series can be differentiated term by term on any such interval which is entirely within the interval of convergence and the sum of a convergent powers series is continuous within such an interval. There is a theorem due to Abel which holds that where a power series converges up to and including an endpoint of its interval of convergence, the interval of uniform convergence also extends so far as to include this endpoint. For detailed

proofs of these propositions see David Bressoud's excellent textbook [2]. The theory of power series is therefore the best starting point for getting confidence that the usual manipulations are actually valid.

If you were pushed to prove that you can actually perform term by term differentiation as in (40) you would need to fall back on proving uniform convergence of the derived series (noting that s^k is continuous and its derivatives are also continuous) and if you want to see how that might be done see the Appendix. In practical terms if we can establish a power series representation with a positive radius of convergence we get all the great properties such as uniform convergence, term by term differentiation etc so that there is no need to perform detailed proofs. The power series representation of e^x is of course the paradigm example of such properties.

In (40) if we let $s = 1$ we formally get that $\sum_k kp_k = E[X]$, the expectation of X. Whenever this expection exists, the derivative $P'(s)$ will be continuous in the closed interval $-1 \leq s \leq 1$. Note here that our series $P(s)$ certainly converges for $-1 < s < 1$ since the coefficients are bounded (they must sum to 1 being probabilities) and we get the nice properties for power series as a result. If $\sum_k kp_k$, diverges then $P'(s) \to \infty$ as $s \to 1$ which is to say that X has infinite expectation which is formally stated as $P'(1) = E[X] = \infty$.

By noting that $P(1) = 1$, equation (39) becomes amenable to the mean value theorem, that is, there is some λ between 1 and s such that $Q(s) = \frac{P(1)-P(s)}{1-s} = P'(1)$. Because both $P'(s)$ and $Q(s)$ are monotone they must have the same finite or infinite limits denoted by $P'(1)$ and $Q(1)$. From this it follows that:

$$\mathbb{E}[X] = \sum_{j=1}^{\infty} jp_j = \sum_{k=0}^{\infty} q_k \tag{41}$$

or in terms of generating functions:

$$\boxed{\mathbb{E}[X] = P'(1) = Q(1)} \tag{42}$$

If we differentiate (40) we get:

$$P''(s) = \sum_{k=1}^{\infty} k(k-1)p_k s^{k-2} \tag{43}$$

so that $P''(1) = \sum_{k=1}^{\infty} k(k-1)p_k$. But differentiating (39) gives:

$$\begin{aligned} P'(s) =& Q(s) - (1-s)Q'(s) \\ P''(s) =& 2Q'(s) - (1-s)Q''(s) \end{aligned} \tag{44}$$

Thus:

$$\mathbb{E}[X(X-1)] = \sum_{k=1}^{\infty} k(k-1)p_k = P''(1) = 2Q'(1) \tag{45}$$

Just recapping what the variance var(X) $= \sigma^2$ is we have:

$$\begin{aligned}\sigma^2(X) =&\mathbb{E}[(X-\mathbb{E}[X])^2]\\ =&\mathbb{E}[X^2-2X\mathbb{E}[X]+\mathbb{E}^2[X]]\\ =&E[X^2]-2\mathbb{E}^2[X]+\mathbb{E}^2[X]\\ =&\mathbb{E}[X^2]-\mathbb{E}^2[X]\end{aligned} \tag{46}$$

Using (42), (44) and (45) we have:

$$\begin{aligned}\sigma^2(X) =&\mathbb{E}[X(X-1)]+\mathbb{E}[X]-\mathbb{E}^2[X]\\ =&P''(1)+P'(1)-P'^2(1)\\ =&2Q'(1)+Q(1)-Q^2(1)\end{aligned} \tag{47}$$

5 The concept of convolution

If we have a random variable X which assumes only non-negative integral values then s^X is a well defined random variable and the generating function approach developed above leads us to conclude that the generating function of the probability distribution of X can be written compactly as:

$$\mathbb{E}[s^X] \tag{48}$$

since this amounts to $p_0s^0+p_1s^1+p_2s^2+\ldots$ (see (34)).

Moreover, if X and Y are independent then so are s^X and s^Y and $\mathbb{E}[s^{X+Y}]=\mathbb{E}[s^X]\,\mathbb{E}[s^Y]$. A quick proof of the general proposition is in the Appendix.

The concept of convolution in the discrete case essentially revolves around multiplication of series and the Cauchy product sum for the general term. So let X and Y be non-negative independent integral valued random variables with respective probability distributions $\mathbb{P}[X=j]=a_j$ and $\mathbb{P}[Y=j]=b_j$. Clearly the event $(X=j,Y=k)$ has probability a_jb_k. Now consider the random variable of the sum $S=X+Y$ and we can construct the event $S=r$ as the union of mutually disjoint (exclusive) events:

$(X=0,Y=r),(X=1,Y=r-1),\ldots,(X=r,Y=0)$

Hence the distribution $c_r=\mathbb{P}[S=r]$ is given by the Cauchy product sum :

$$\boxed{c_r=a_0b_r+a_1b_{r-1}+a_2b_{r-2}+\cdots+a_{r-1}b_1+a_rb_0=\sum_{k=0}^{r}a_kb_{r-k}} \tag{49}$$

Note that (49) just reflects the addition axiom of mutually exclusive events the components of which are products of the individual probabilities arising from independent events.

This in essence is the convolution and is written:

$$\boxed{\{c_k\} = \{a_k\} * \{b_k\}} \tag{50}$$

This formulation is quite general and the a_k and b_k don't necessarily need to be probability distributions. Equation (50) is meant to convey the fact that $c_r = \sum_{k=0}^{r} a_k b_{r-k}$.

Now because the sequences $\{a_k\}$ and $\{b_k\}$ have generating functions $A(s) = \sum_k a_k s^k$ and $B(s) = \sum_k b_k s^k$ respectively, the product $A(s)B(s)$ is just the product of the two power series ie

$$\begin{aligned} A(s)B(s) =& (a_0 + a_1 s + a_2 s^2 + a_3 s^3 + \dots)\,(b_0 + b_1 s + b_2 s^2 + b_3 s^3 + \dots) \\ =& a_0 b_0 + (a_0 b_1 + a_1 b_0) s + (a_0 b_2 + a_1 b_1 + a_2 b_0) s^2 + \dots \\ =& \sum_{r=0}^{\infty} c_r s^r \end{aligned} \tag{51}$$

where $c_r = \sum_{k=0}^{r} a_k b_{r-k}$

Thus we have the result that if $\{c_k\}$ is the convolution of $\{a_k\}$ and $\{b_k\}$, then the generating function $C(s) = \sum_k c_k s^k$ is simply:

$$\boxed{C(s) = A(s)B(s)} \tag{52}$$

If we assume that X and Y are non-negative integral valued mutually independent random variables with generating functions $A(s)$ and $B(s)$, then their sum $X+Y$ has generating function $A(s)B(s)$, which justifies the comments made following (48).

It is clear from this development that we have commutativity and associativity. For instance, $\{a_k\} * \{b_k\} = \{b_k\} * \{a_k\}$ and $\{a_k\} * \{b_k\} * \{c_k\} = \{c_k\} * \{b_k\} * \{a_k\}$.

If we have the sum of n independent random variables X_i which have the same common probability distribution $\{a_k\}$ then the distribution of the sum $S_n = X_1 + X_2 + \cdots + X_n$ is $\underbrace{\{a_k\} * \{a_k\} * \cdots * \{a_k\}}_{\text{n terms}} = \{a_k\}^{n*} = \{a_k\}^{(n-1)*} * \{a_k\}$.

The sequence of numbers $\{a_k\}^{n*}$ has the generating function $A^n(s)$.

In the continuous context, if X and Y are two independent continuous random variables and $Z = X + Y$ then the probability density of Z, $f_Z(z)$ is:

$$f_Z(z) = f_{X+Y}(z) = \int_{-\infty}^{\infty} f_Y(z-x) f_X(x)\,dx = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\,dy \tag{53}$$

Note the discrete analogue with $c_r = \sum_{k=0}^{r} a_k b_{r-k}$ in (48).

Convolution is fundamental in Fourier theory and is defined as follows [see [4] pages 139-140]:

$$\boxed{(f * g)(x) = \int_{-\infty}^{\infty} f(x-t)g(t)\,dt} \tag{54}$$

If f, g inhabit Schwartz space then convergence of the integral is assured and the convolution itself lives in Schwartz space (see [4] page 142) and the big relationship is that the Fourier transform of the convolution is the product of the Fourier transforms of f and g which echoes (52). The other big idea from Fourier theory is that the Fourier transform of a Gaussian is a Guassian Indeed, the Gaussian is a fixed point in Schwartz space. The Laplace transform of a Gaussian is not a Gaussian (see `https://www.gotohaggstrom.com/The%20Laplace%20transform%20of%20a%20GaussianV3.pdf`). For all the details on the basics of Fourier theory see [3].

6 Compound distributions

If $\{X_i\}$ is a sequence of independent integral valued random variables with common probability distribution $\mathbb{P}[X_i = k] = p_k$ then the variable $Y = \sum_{k=1}^{n} X_k$ where n is an integral valued random variable, independent of all of the X_k, is said to be a compound distribution. Clearly, Y is also an integral valued random variable.

Let the probability function for n be given by $\mathbb{P}[n = k] = g_k$ for $k \geq 0$ and let its probability generating function be $G(s) = \sum_{k=0}^{\infty} g_k s^k$. Let the probability generating function for each X_k be $P(s) = \sum_{k=0}^{\infty} p_k s^k$.

Using basic addition and multiplication axioms we have that:

$$\begin{aligned} \mathbb{P}[Y = j] &= \sum_{k=0}^{\infty} \mathbb{P}[n = k]\, \mathbb{P}\Big[\sum_{i=1}^{k} X_i = j\Big] \\ &= \sum_{k=0}^{\infty} g_k\, \mathbb{P}\Big[\sum_{i=1}^{k} X_i = j\Big] \end{aligned} \tag{55}$$

Now if $Y(s)$ is the probability generating function for Y, the coefficient of s^j in $Y(s)$ must be $\sum_{k=0}^{\infty} g_k\, \mathbb{P}[\sum_{i=1}^{k} X_i = j]$. Since the X_i are independent with common probability generating function $P(s)$, the probability that $\sum_{i=1}^{k} X_i = j$ for given k, is the coefficient of s^j in $[P(s)]^k$. Hence $\mathbb{P}[Y = j]$ = coefficient of s^j in $\sum_{k=0}^{\infty} g_k\, [P(s)]^k$. This means that:

$$Y(s) = \sum_{k=0}^{\infty} g_k\, [P(s)]^k \tag{56}$$

But given the way $G(s)$ is defined above we must then have:

$$\boxed{Y(s) = G[P(s)]} \tag{57}$$

It is now possible to obtain the mean and variance of Y without actually obtaining an explicit expression for its probability function. First of all we note that since $P(s) = \sum_{k=0}^{\infty} p_k\, s^k$ we have that $P(1) = \sum_{k=0}^{\infty} p_k = 1$. Hence the mean of Y is, using (42) and (57):

$$\begin{aligned}\mathbb{E}[Y] =& Y'(1)\\ =& G'[P(1)]\, P'(1)\\ =& G'(1)\, P'(1)\\ =& \mathbb{E}[n]\, \mathbb{E}[X]\end{aligned} \tag{58}$$

This can also be written in the usual "mu" notations as $\mu_Y = \mu_n\, \mu_x$.

The variance of Y is worked out as follows using (47) and noting that $P(1) = 1$:

$$\begin{aligned}\text{var(Y)} =& Y''(1) + Y'(1) - [Y'(1)]^2\\ =& G''[P(1)]\, [P'(1)]^2 + G'[P(1)]\, P''(1) + G'(1)\, P'(1) - [G'(1)]^2\, [P'(1)]^2\\ =& \{G''(1) - [G'(1)]^2\}\, [P'(1)]^2 + G'(1)\{P''(1) + P'(1)\}\\ =& \Big\{\text{Var(n)} \ - \mathbb{E}[n]\Big\}[\mathbb{E}[X]]^2 + \mathbb{E}[n]\Big\{\text{var(X)} \ + [\mathbb{E}[X]]^2\Big\}\\ =& [\mathbb{E}[X]]^2 \text{var(n)} + \mathbb{E}[n] \text{var(X)}\end{aligned} \tag{59}$$

6.1 Some fundamental examples of probability generating functions

Binomial distribution

If we let $b(k, n, p)$ be the probability that n Bernoulli trials with probability p for success and $q = 1 - p$ for failure result in k successes and $n - k$ failures we have that:

$$b(k, n, p) = \binom{n}{k} p^k q^{n-k} \tag{60}$$

The generating function for this binomial distribution is:

$$B(s) = \sum_{k=0}^{n} \binom{n}{k} (ps)^k q^{n-k} = (q + ps)^n \tag{61}$$

Compare this to the moment generating function in (10).

This is because we have the sum of n indepedent random variables with the common generating function $q + ps$ ie each variable X_k assumes the value 0 with probability q and the value 1 with probability p. We can write $\{b(k, n, p)\} = \{b(k, 1, p)\}^{n*}$. Because:

$$(q + ps)^m\, (q + ps)^n = (q + ps)^{m+n} \tag{62}$$

we must have that:

$$\{b(k,m,p)\} * \{b(k,n,p)\} = \{b(k,m+n,p)\} \tag{63}$$

Now recalling (42), if we differentiate (61) we get:

$$B'(s) = np(q+ps)^{n-1} \tag{64}$$

and hence:

$$\mathbb{E}[S_n] = B'(1) = np(q+p) = np \tag{65}$$

since $p+q=1$.

We get the variance from equation (47) noting that $P''(s) = n(n-1)p^2(q+ps)^{n-2}$:

$$\begin{aligned} var(S_n) =& B''(1) + B'(1) - B'^2(1) \\ =& n(n-1)p^2 + np - n^2p^2 \\ =& np(1-p) \\ =& npq \end{aligned} \tag{66}$$

Poisson distribution

The Poisson distribution $\mathbb{P}$oi[k, λ] with parameter $\lambda > 0$ has the form:

$$\mathbb{P}\text{oi}[k,\lambda] = \frac{e^{-\lambda}\lambda^k}{k!} \tag{67}$$

for $k = 0, 1, 2, \ldots$

The generating function is:

$$\text{Poi}_k(s) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{(\lambda s)^k}{k!} = e^{-\lambda+\lambda s} \tag{68}$$

Compare this to the moment generating function in (21).

6.2 An integral valued problem

Four tickets are drawn, one at a time with replacement, from a set of ten tickets numbered $1, 2, \ldots, 10$ in such a way that at each draw each ticket is equally likely to be selected. What is the probability that the total of the numbers on the four drawn tickets is 20?

Solution

This has been set up so we have independence and equal probabilities. It is the type of problem they set in actuarial courses. Thus if x_i denotes the number on the i^{th} ticket then

for $i = 1, 2, 3, 4$, x_i is an integral valued variate with possible values $1, 2, 3, \ldots, 10$ each having associated probability $\frac{1}{10}$. Hence each x_i has probability generating function:

$$\frac{1}{10}s + \frac{1}{10}s^2 + \frac{1}{10}s^3 + \cdots + \frac{1}{10}s^{10} = \frac{s}{10}(1 + s + s^2 + \cdots + s^9) = \frac{s}{10}(1 - s^{10})(1 - s)^{-1} \tag{69}$$

Since the x_i are clearly independent it follows that the total of the numbers on the drawn tickets, namely, $z = x_1 + x_2 + x_3 + x_4$ has probabiity generating function:

$$\left[\frac{s}{10}(1 - s^{10})(1 - s)^{-1}\right]^4 = \frac{1}{10^4}s^4(1 - s^{10})^4(1 - s)^{-4} \tag{70}$$

The required probability is the coefficient of s^{20} in (70) which is equal to the coefficient pf s^{16} in the following expression:

$$\frac{1}{10^4}(1 - s^{10})^4(1 - s)^{-4} \tag{71}$$

We need to expand this expression and the only term that presents any difficulty is $(1 - s)^{-4}$. If you have forgotten the general form $(1 - x)^{-n}$ see `https://gotohaggstrom.com/The%20binomial%20series%20for%20negative%20integral%20exponents.pdf`.

We have that:

$$(1-x)^{-n} = 1 + nx + \frac{n(n+1)}{1.2}x^2 + \frac{n(n+1)(n+2)}{1.2.3}x^+ \cdots + \frac{n(n+(n+2)\ldots(n+k-1)}{1.23\ldots.k}x^k + \ldots \tag{72}$$

Expanding (71) we have:

$$\begin{aligned}
\frac{1}{10^4}(1 - s^{10})^4(1 - s)^{-4} =& \frac{1}{10^4}\left(1 - 4s^{10} + 6s^{20} - \ldots\right)\left(1 + 4s + \frac{4.5}{1.2}s^2 + \frac{4.5.6}{1.2.3}s^3 + \right. \\
& \frac{4.5.6.7}{1.2.3.4}s^4 + \frac{4.5.6.7.8}{1.2.3.4.5}s^5 + \frac{4.5.6.7.8.9}{1.2.3.4.5.6}s^6 + \cdots + \\
& \left.\frac{4.5.6.7.8.9.10.11.12.13.14.15.16.17.18.19}{16!}s^{16} + \ldots\right) \\
=& \frac{1}{10^4}\left(-4\frac{4.5.6.7.8.9}{1.2.3.4.5.6} + \frac{4.5.6.7.8..10.11.12.13.14.15.16.17.18.19}{16!}s^{16}\right) \\
=& \frac{1}{10^4}\left(-4\frac{7.8.9}{1.2.3} + \frac{17.18.19}{1.2.3}\right) \\
=& \frac{633}{10,000}
\end{aligned} \tag{73}$$

This would clearly be an horrendous calculation to perform by enumeration of cases, but by reducing the dimension of the problem one can check by hand that it actually gives the correct number of cases, as it should because the multiplication generates all the relevant terms. So let's assume we are now looking for $z = x_1 + x_2 + x_3 = 15$. As before we are after the coefficient of s^{15} in:

$$\frac{s^3}{10^3}(1-s^{10})^3(1-s)^{-3} \tag{74}$$

which amounts to the coefficient of s^{12} in:

$$\begin{aligned}\frac{1}{1000}(1-s^{10})^3(1-s)^{-3} =& \frac{1}{1000}\left(1-3s^{10}+\dots\right)\left(1+3s+\frac{3.4}{1.2}s^2+\dots+\frac{3.4.5....13.14}{12!}s^{12}\right)\\ =& \frac{1}{1000}\left(1-3s^{10}+\dots\right)\left(1+3s+\frac{3.4}{1.2}s^2+\dots+\frac{13.14}{1.2}s^{12}\right)\end{aligned} \tag{75}$$

which is:

$$\frac{1}{1000}\left(-3\frac{3.4}{1.2}+\frac{13.14}{1.2}\right) = \frac{-18+91}{1000} = \frac{73}{1000} \tag{76}$$

We can manually confirm this number by exhaustively working out the relevant triples summing to 15 as follows (noting that for abc there are $3! = 6$ cases and for aab there are $\frac{3!}{2} = 3$ cases:

Triple	**Number of cases**
$10, 4, 1$	6
$10, 3, 2$	6
$9, 5, 1$	6
$9, 4, 2$	6
$9, 3, 3$	3
$8, 6, 1$	6
$8, 5, 2$	6
$8, 4, 3$	6
$7, 7, 1$	3
$7, 6, 2$	6
$7, 5, 3$	6
$7, 4, 4$	3
$6, 6, 3$	3
$6, 5, 4$	6
$5, 5, 5$	1
Total cases	73

(77)

6.3 The relationship between generating functions and moment generating functions for positive integral valued variables

The relationship between a generating function and a moment generating function is developed by Feller in a problem (see [3], page 285, Problem 24) which is as follows:

" Let $\mathbf{X}$ be a random variable with generating function $P(s)$, and suppose that $\sum p_n s^n$ converges for some $s_0 > 1$. Then all the moments $m_r = \mathbb{E}[\mathbf{X}^r]$ exist and the generating function $F(s)$ of the sequence $\frac{m_r}{r!}$ converges at least for $|s| < \ln s_0$. Moreover:

$$F(s) = \sum_{r=0}^{\infty} \frac{m_r}{r!} s^r = P(e^s) \tag{78}$$

Feller notes that $F(s)$ is called the moment generating function, although in reality it generates $\frac{m_r}{r!}$ "

Thus we have that:

$$P(s) = \sum_{k=0}^{\infty} p_k s^k \tag{79}$$

and we know that there is some $s_0 > 1$. such that $P(s)$ converges. It actually converges absolutely since the terms are all positive. Because we are dealing with positive integral valued variables our moments for fixed r become:

$$\mathbb{E}[X^r] = \sum_{k=0} p_k \, k^r \tag{80}$$

We have to prove that for each fixed r this series converges to a finite number. The crux of the proof is the observation that because s_0^k will eventually dominate k^r we can dominate the moment sum by something convergent. In essence there is some k' such that $s_0^k > k^r$ for all $k > k'$. We can in principle demonstrate this k' by noting that:

$s_0^k - k^r = e^{k \ln s_0} - e^{r \ln k} > 0$ if $\frac{k}{\ln k} > \frac{r}{\ln s_0}$. So k' has to satisfy that inequality. As a numerical example, suppose that $s_0 = 2$ and $r = 5$. We then need $\frac{k}{\ln k} > \frac{5}{\ln 2} = 7.21$ and so $k' = 23$. Hence for all $k \geq 23$ we will have $2^k > k^5$.

Thus we have:

$$\begin{aligned}
\sum_{k=0}^{\infty} p_k k^r &= \sum_{k=0}^{k'-1} p_k k^r + \sum_{k=k'}^{\infty} p_k k^r \\
&= A + \sum_{k=k'}^{\infty} p_k k^r \\
&\leq A + \sum_{k=k'}^{\infty} p_k s_0^r \\
&< A + B \\
&< \infty
\end{aligned} \tag{81}$$

Where A is some positive number and $B = P(s_0)$ and so $\sum_{k=k'}^{\infty} p_k s_0^r$ is certainly less than B.

Thus all the moments exist as finite numbers. To get (78) we can proceed as follws:

$$
\begin{aligned}
P(e^s) &= \sum_{k=0}^{\infty} p_k (e^s)^k \\
&= \sum_{k=0}^{\infty} p_k e^{sk} \\
&= \sum_{k=0}^{\infty} p_k \Big(\sum_{r=0}^{\infty} \frac{(sk)^r}{r!} \Big) \\
&= \sum_{k=0}^{\infty} p_k \Big(\sum_{r=0}^{\infty} \frac{s^r k^r}{r!} \Big) \\
&= \sum_{r=0}^{\infty} \Big(\sum_{k=0}^{\infty} p_k k^r \Big) \frac{s^r}{r!} \\
&= \sum_{r=0}^{\infty} m_r \frac{s^r}{r!} \\
&= F(s)
\end{aligned}
\tag{82}
$$

provided $|s| < \ln s_0$ for then $s_0 > e^{|s|} > 1$ which was the condition that ensures the assumed convergence of $P(s)$. Note that the rearrangement of the series is justified by one of Dirichlet's theorems which holds that you can rearrange the terms of an absolutely convergent series.

Typically the relationship described above is arrived at using another route which assumes all the nice properties. For finite n we know that the expectation operator distributes as follows:

$$
\mathbb{E}[\sum_{i=1}^{n} a_i X_i] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i] \tag{83}
$$

It is then assumed that this process holds for infinite n and that one can also perform differentiation an arbitrary number of times. Because e^x is the paradigm case of a power series with an infinite radius of convergence which is uniformly continuous on $\mathbb{R}$ and also infinitely differentiable, we can write:

$$
e^{sX} = \sum_{r=0}^{\infty} \frac{(sX)^r}{r!} \tag{84}
$$

Now if $M(s)$ exists for some $s \in (-h, h)$ for some $h > 0$ we have:

$$
M(s) = \mathbb{E}[e^{sX}] = \mathbb{E}\Big[\sum_{r=0}^{\infty} \frac{(sX)^r}{r!} \Big] \tag{85}
$$

We then interchange the expectation operator with the summation and get:

$$
\boxed{M(s) = \sum_{r=0}^{\infty} \mathbb{E}[X^r] \frac{s^r}{r!}} \tag{86}
$$

This means that the coefficient of $\frac{s^r}{r!}$ in the power series expansion of $M(s)$ is $\mathbb{E}[X^r]$.

But the Maclaurin series expansion for $M(s)$ is:

$$M(s) = \sum_{r=0}^{\infty} M^{(r)}(0)\frac{s^r}{r!} \tag{87}$$

where $M^{(r)}(0) = \frac{d^r}{ds^r}M(s)\Big|_{s=0}$.

So when we compare coefficients of s^r in the two power series representations for $M(s)$ we see that:

$$\boxed{\mathbb{E}[X^r] = M^{(r)}(0)} \tag{88}$$

What this gives us is two ways of finding moments:

(1) If you can expand the mgf as a power series of s then the coefficient of s^r multiplied by $r!$ gives $\mathbb{E}[X^r]$; or

(2) if the mgf can be differentiated infinitely then the r^{th} order derivative evaluated at $s = 0$ is $\mathbb{E}[X^r]$

In relation to (2) we can proceed formally as follows:

$$M^{(r)}(s) = \frac{d^r}{ds^r}\mathbb{E}[e^{sX}] = \mathbb{E}\Big(\frac{d^r}{ds^r}e^{sX}\Big) = \mathbb{E}[X^r e^{sX}] \tag{89}$$

so on setting $s = 0$ we have $\mathbb{E}[X^r] = M^{(r)}(0)$.

We can try this approach out on some examples.

Binomial example

We know from (10) that the mgf of a random variable X which is $B(n, p)$ is $M(s) = [pe^s + (1 - p)]^n$. If, for instance, we want $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ we work out the first and second derivatives as follows:

$$\mathbb{E}[X] = \frac{d}{ds}M(s)\Big|_{s=0} = n[pe^s + (1 - p)]^{n-1}\, pe^s\Big|_{s=0} = np \tag{90}$$

and

$$\mathbb{E}[X^2] = \frac{d^2}{ds^2}M(s)\Big|_{s=0} = n(n-1)[pe^s+(1-p)]^{n-2}\, p^2e^{2s}\Big|_{s=0} + n[pe^s+(1-p)]^{n-1}\, pe^s\Big|_{s=0} = n(n-1)p^2+np \tag{91}$$

This is a lot easier than expanding as a power series.

Poisson example

As another example consider the mgf for the Poisson distribution with parameter λ given in (21) ie $M(s) = e^{\lambda(e^s-1)}$. To find $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ we proceed as follows:

$$\mathbb{E}[X] = \frac{d}{ds}M(s)\Big|_{s=0} = e^{\lambda(e^s-1)}\lambda e^s\Big|_{s=0} = \lambda \tag{92}$$

and

$$\mathbb{E}[X^2] = \frac{d^2}{ds^2}M(s)\Big|_{s=0} = e^{\lambda(e^s-1)}\lambda e^s + e^{\lambda(e^s-1)}\lambda^2 e^{2s}\Big|_{s=0} = \lambda^2 + \lambda \tag{93}$$

Normal example

Let's suppose that X is normally distributed with mean 0 and variance σ^2. As noted in (29) the mgf for an $N(\mu, \sigma^2)$ normal distribution is:

$$M(s) = e^{\mu s + \frac{\sigma^2 s^2}{2}} \tag{94}$$

so for an $N(0, \sigma^2)$ distribution we have:

$$\begin{aligned} M(s) =& e^{\frac{\sigma^2 s^2}{2}} \\ =& \sum_{r=0}^{\infty} \Big(\frac{\sigma^2 s^2}{2}\Big)^r \frac{1}{r!} \\ =& \sum_{r=0}^{\infty} \frac{\sigma^{2r}}{2^r} \frac{s^{2r}}{r!} \\ =& \sum_{r=0}^{\infty} \frac{\sigma^{2r}(2r)!}{2^r r!} \frac{s^{2r}}{(2r)!} \end{aligned} \tag{95}$$

Clearly the coefficient of s^{2r+1} is zero for every non-negative integral r and therefore $\mathbb{E}[X^{2r+1}] = 0$. We can see that the coefficient of $\frac{s^{2r}}{(2r)!}$ is $\frac{\sigma^{2r}(2r)!}{2^r r!}$ for $r = 0, 1, 2, \ldots$. In summary:

$$\boxed{\begin{array}{cc} \mathbb{E}[X^{2r+1}] = 0 & r = 0, 1, 2, \ldots \\ \mathbb{E}[X^{2r}] = \frac{\sigma^{2r}(2r)!}{2^r r!} & r = 0, 1, 2, \ldots \end{array}} \tag{96}$$

The fact that odd moments are zero follows from the fact that the normal distribution is symmetric about 0.

On the other hand if we differentiate we get:

$$\frac{d}{ds}M(s) = \sigma^2 s e^{\frac{\sigma^2 s^2}{2}} \tag{97}$$

Thus $\mathbb{E}[X] = M^{(1)}(0) = 0$

and

$$\frac{d^2}{ds^2}M(s) = (\sigma^2 + \sigma^4\, s^2) e^{\frac{\sigma^2 s^2}{2}} \tag{98}$$

So $\mathbb{E}[X^2] = M^{(2)}(0) = \sigma^2$.

You will note that it is hard to see the pattern for an arbitrary r using this method.

A continuous random variable with distribution function F

Let X be any continuous random variable having the distribution function F and pdf f. Let's work out the distribution of $Y = F(X)$ over the interval $[0, 1]$.

The mgf of Y is:

$$M_Y(s) = \mathbb{E}[e^{sY}] = \mathbb{E}[e^{sF(x)}] = \int_{-\infty}^{\infty} e^{sF(x)} f(x)\, dx \tag{99}$$

Recall that $F(x) = \int_{-\infty}^{x} f(t)\, dt$ and hence $dF(x) = f(x)dx$ and so for $s \neq 0$ we have:

$$\int_{-\infty}^{\infty} e^{sF(x)}\, dx = \frac{e^{sF(x)}}{s}\Big|_{-\infty}^{\infty} = \frac{e^s - 1}{s} \tag{100}$$

noting that $F(\infty) = 1$ and $F(-\infty) = 0$ and $M_Y(0) = \mathbb{E}[e^{0Y}] = 1$.

But from (23), (100) is recognisable as a uniform distribution on $[0, 1]$.

Recognizing distributions from their mgfs

After some practice one can simply look at an mgf and read off the distribution. Here are some examples:

$$M(s) = \sum_{r=0}^{\infty} \frac{s^{2r}}{r!} \tag{101}$$

This is $N(0, 2)$ because the mgf for $N(0, \sigma^2)$ is :

$$M(s) = \sum_{r=0}^{\infty} \left(\frac{\sigma^2 s^2}{2}\right)^r \frac{1}{r!} \tag{102}$$

so if $\sigma^2 = 2$ we get our result.

Suppose now that the mgf is:

$$M(s) = \left[\frac{1}{4} + \frac{3}{4}\sum_{r=0}^{\infty} \frac{s^r}{r!}\right]^{10} \tag{103}$$

Once you realise that $e^s = \sum_{r=0}^{\infty} \frac{s^r}{r!}$ it is all over since the mgf is clearly of a binomial character. In fact it is $B(10, \frac{3}{4})$ because $M(s) = \left[\frac{1}{4} + \frac{3}{4}e^s\right]^{10}$ where $p = \frac{3}{4}$. See (10)

The uniqueness property guarantees that the mgf correctly "fingers" the distribution.

Recall that we can write a generating function for positive integral values of a random variable X as $\mathbb{E}[s^X]$. This formulation is also called a factorial moment generating function ("fmgf") for reasons to be given shortly. The fmfg is also referred to as a probability generating function. If we start with the basic relationship $g(s) = \mathbb{E}[s^X]$ we perform the following manipulation:

$$g(s) = \mathbb{E}[s^X] = \mathbb{E}[e^{X \ln s}] = M(\ln s) \text{ see (4)} \tag{104}$$

and

$$M(s) = \mathbb{E}[e^{sX}] = \mathbb{E}[\left(e^s\right)^X] = g(e^s) \tag{105}$$

What this means is that the mgf evaluated at $\ln s$ gives the fmgf and the fmgf evaluation at e^s gives the mgf:

$$\boxed{\text{mgf} \xrightarrow{\ln s} \quad \text{fmgf}} \tag{106}$$

$$\boxed{\text{fmgf} \xrightarrow{e^s} \quad \text{mgf}} \tag{107}$$

If the mgf involves s only through e^s then all that is needed is to replace e^s by s in order to obtain the fmgf from the mgf and if the fmgf involves s only through powers of s then replace s by e^s to obtain the mgf from the fmgf.

If X is $B(n,p)$ we know its mgf is $M(s) = [pe^s + (1-p)]^n$ so we can immediately see that the fmgf is given by $g(s) = [ps + (1-p)]^n$ (simply replace e^s by s which is equivalent to $M(\ln s) = [pe^{\ln s} + (1-p)]^n = [ps + (1-p)]^n$).

Why "factorial" moment generating function?

Assuming that we can perform repeated differentiation of $g(s) = \mathbb{E}[s^X]$ we have:

$$g^{(r)}(s) = \mathbb{E}[X(X-1)\dots(X-r+1)s^{X-r}] \tag{108}$$

Hence when we evaluate at $s = 1$ we have:

$$g^{(r)}(1) = \mathbb{E}[X(X-1)\dots(X-r+1)] \tag{109}$$

Applying this concept to a $B(n,p)$ random variable with fmgf $g(s) = [ps + (1-p)]^n$ we get the following:

$$\begin{aligned} g^{(1)}(s) =& np[ps + (1-p)]^{n-1} \\ g^{(2)}(s) =& n(n-1)p^2[ps + (1-p)]^{n-2} \\ &\vdots \\ g^{(r)}(s) =& n(n-1)\dots(n-r+1)p^r[ps + (1-p)]^{n-r} \text{ for any integer } r \text{ with } 1 \le r \le n \end{aligned} \tag{110}$$

If we evaluate these derivatives at $s = 1$ we have:

$$\begin{aligned} \mathbb{E}[X] = g^{(1)}(1) =& np \\ \mathbb{E}[X(X-1)] = g^{(2)}(1) =& n(n-1)p^2 \\ &\vdots \\ \mathbb{E}[X(X-1)\dots(X-r+1)] = g^{(r)}(1) =& n(n-1)\dots(n-r+1)p^r \end{aligned} \tag{111}$$

The probabilities are generated in accordance with this formula:

$$\boxed{\mathbb{P}[X = r] = \frac{g^{(r)}(0)}{r!} \quad r = 0, 1, 2, \dots} \tag{112}$$

where since $g(0) = \mathbb{P}[X = 0]$ we interpret $g^{(0)}(0)$ as equal to $g(0)$. To establish this result we let

$\mathbb{P}[X = i] = p_i$ for $i = 0, 1, 2, \ldots$ so we get:

$$g(s) = \mathbb{E}[s^X] = \sum_{i=0}^{\infty} p_i s^i \tag{113}$$

Because the p_i are probabilities with $0 \leq p_i < 1$, $g(s)$ is dominated by a convergent geometric series so that we can differentiate an arbitrary number of times for $s \in (-1, 1)$. Thus:

$$g^{(r)}(s) = \sum_{i=r}^{\infty} i(i-1)\ldots(i-r+1)s^{i-r}\, p_i \tag{114}$$

Clearly if we set $s = 0$ the RHS of (114) reduces to the term for which $i = r$ which is just a constant term:

$$g^{(r)}(0) = r(r-1)\ldots 2.1 = r!p_r \tag{115}$$

and thus we get (112).

7 Uniqueness Theorem

If two random variables for the same moment generating functions (mgf's), then they have the same distribution, and conversely.

This extremely powerful result is difficult to prove in detail and a thumbnail sketch of the proof is given in the Appendix.

Let's suppose we have two random variables X and Y, then $M_X(s) = M_Y(s)$ for $s \in (-h, h)$ for some $h > 0$ if and only if $F_X(u) = F_Y(u)$ for all real u. Thus if you do a problem and arrive at $M(s) = e^{2(e^s-1)}$ then you can conclude with confidence that X has a Poisson distribution with parameter $\lambda = 2$ and no other distribution can have such a distribution - see (21).

In the most general setting we will have random variables $X_1, X_2, \ldots, X_n$ with given density $f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$ and given functions $g_1(.,.,\ldots.), \ldots, g_k(.,.,\ldots)$ and you have to find the joint distribution of $Y_1 = g_1(X_1, X_2, \ldots, X_n), \ldots, Y_k = g_1(X_1, X_2, \ldots, X_n)$. Assuming the joint moment generating function exists it will take the form:

$$\begin{aligned} M_{Y_1,\ldots,Y_k}(s_1, \ldots, s_k) =& \mathbb{E}[e^{s_1 Y_1 + \cdots + s_k Y_k}] \\ =& \boxed{\int \cdots \int e^{s_1 g_1(x_1,\ldots,x_n) + \cdots + s_k g_k(x_1,\ldots,x_n)} \times f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) \prod_{i=1}^{n} dx_i} \end{aligned} \tag{116}$$

For $k > 1$ this technique is hard because we can only recognise a small number of joint moment generating functions.

Example
Suppose X and Y are independent random variables with the following probability functions:

$\mathbb{P}[X = x] = \frac{x}{6}$ for $x = 1, 2, 3$

and

$\mathbb{P}[Y = y] = \frac{y+2}{10}$ for $y = -1, 2, 3$

If we want the distributions of $X + Y$ and $X - Y$ we can use mgfs as follows. The relevant mgfs are:

$$M_X(s) = \frac{1}{6}e^s + \frac{2}{6}e^{2s} + \frac{3}{6}e^{3s} \tag{117}$$

and

$$M_Y(s) = \frac{1}{10}e^{-1.s} + \frac{4}{10}e^{2s} + \frac{5}{10}e^{3s} \tag{118}$$

Since X and Y are independent we know that $M_{X+Y}(s) = M_X(s)M_Y(s)$ which gives:

$$\begin{aligned} M_{X+Y}(s) =& \left(\frac{1}{6}e^s + \frac{2}{6}e^{2s} + \frac{3}{6}e^{3s}\right)\left(\frac{1}{10}e^{-1.s} + \frac{4}{10}e^{2s} + \frac{5}{10}e^{3s}\right) \\ =& \frac{1}{60}e^{0.s} + \frac{4}{60}e^{3s} + \frac{5}{60}e^{4s} + \frac{2}{60}e^s + \frac{8}{60}e^{4s} + \frac{10}{60}e^{5s} + \frac{3}{60}e^{2s} + \frac{12}{60}e^{5s} + \frac{15}{60}e^{6s} \\ =& \frac{1}{60}e^{0.s} + \frac{2}{60}e^s + \frac{3}{60}e^{2s} + \frac{4}{60}e^{3s} + \frac{13}{60}e^{4s} + \frac{22}{60}e^{5s} + \frac{15}{60}e^{6s} \end{aligned} \tag{119}$$

We can conclude from this that $X + Y$ assumes the values $0, 1, 2, 3, 4, 5, 6$ with respective probabilities $\frac{1}{60}, \frac{2}{60}, \frac{3}{60}, \frac{4}{60}, \frac{13}{60}, \frac{22}{60}, \frac{15}{60}$. Note that the proabilities sum to 1.

To work out the mgf of $X - Y$ we proceed as follows, noting that $X - Y = X + (-1)Y$ so that $M_{X-Y}(s) = M_X(s).M_Y(-s)$ see (30).

$$\begin{aligned} M_{X-Y}(s) =& \left(\frac{1}{6}e^s + \frac{2}{6}e^{2s} + \frac{3}{6}e^{3s}\right)\left(\frac{1}{10}e^s + \frac{4}{10}e^{-2s} + \frac{5}{10}e^{-3s}\right) \\ =& \frac{5}{60}e^{-2s} + \frac{14}{60}e^{-s} + \frac{23}{60}e^{0.s} + \frac{12}{60}e^s + \frac{1}{60}e^{2s} + \frac{2}{60}e^{3s} + \frac{3}{60}e^{4s} \end{aligned} \tag{120}$$

Note again that the probabilities sum to 1. We can conclude from this that $X - Y$ assumes the values $-2, -1, 0, 1, 2, 3, 4$ with respective probabilities $\frac{1}{12}, \frac{7}{30}, \frac{23}{60}, \frac{1}{5}, \frac{1}{60}, \frac{1}{30}, \frac{1}{20}$.

Showing that if X is $N(0,1)$ then the distribution of X^2 is chi-square with 1 degree of freedom

As background, in the gamma distribution in (32) if we let $p = 1$ we get the exponential distribution as a special case. Moreover, if we let $\lambda = \frac{1}{2}$ and $p = \frac{n}{2}$ we get the chi-square distribution with n degrees of freedom with mgf:

$$M(s) = \frac{1}{(1-2s)^{\frac{n}{2}}} \text{ , if } s < \frac{1}{2} \tag{121}$$

Let $Y = X^2$. Then the mgf of Y:

$$\begin{aligned} M_Y(s) =& \mathbb{E}[e^{sX^2}] \\ =& \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{sx^2} e^{-\frac{x^2}{2}} \, dx \\ =& \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(1-2s)(\frac{x^2}{2})} \, dx \\ =& \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} \, \frac{du}{(1-2s)^{\frac{1}{2}}} \\ =& \frac{1}{(1-2s)^{\frac{1}{2}}} \end{aligned} \tag{122}$$

where the substitution $(1-2s)^{\frac{1}{2}} x = u$ has been used. The result is a chi-square with 1 degree of freedom ($n = 1$).

7.1 Masters of Biostatistics level problems

The following are some problems from a Masters of Biostatistics course which give a feel for the sort of level of knowledge one needs:

Problem 1

Suppose that X has a uniform distribution on $(a, 1)$ where $0 < a < 1$. Suppose $Q = -\frac{1}{2} \ln \left(\frac{X-a}{1-a} \right)$. How do we find the probability density function of Q using the technique of moment generating functions? Check your result by doing a simulation with $a = 0.3$.

Solution

This is a case in one variable so the moment generating function will hopefully reveal an expression which has a recognisable form. Note that the cumulative distribution function of X is $F_X(x) = \frac{x-a}{1-a}$. The moment generating function of Q is:

$$
\begin{aligned}
M_Q(s) =& \mathbb{E}[e^{sQ}] \\
=& \int_a^1 e^{\frac{-s}{2}\ln\left(\frac{x-a}{1-a}\right)}\,dx \\
=& \int_a^1 e^{\ln\left(\frac{x-a}{1-a}\right)^{-\frac{s}{2}}}\,dx \\
=& \int_a^1 \left(\frac{x-a}{1-a}\right)^{-\frac{s}{2}}\,dx \qquad \text{substitute } u=\frac{x-a}{1-a} \\
=& (1-a)\int_0^1 u^{-\frac{s}{2}}\,du \\
=& (1-a)\Big[\frac{u^{1-\frac{s}{2}}}{1-\frac{s}{2}}\Big]_{u=0}^{u=1} \\
=& \frac{2(1-a)}{2-s}
\end{aligned}
\tag{123}
$$

We know from (33) that $\frac{\lambda}{\lambda-s}$ is the mgf of an exponential distribution with parameter $\lambda=2$ and this is weighted by $(1-a)$. The pdf for an exponential function with parameter $\lambda=2$ is $f(x)=2e^{-2x}$ for $x>0$ and 0 elsewhere . Note that the pdf represented by the mgf $\frac{2(1-a)}{2-s}$ must be weighted by $\frac{1}{1-a}$ for the pdf to integrate to 1. That is:

$$
\frac{1}{1-a}\int_0^\infty (1-a)2e^{-2x}\,dx = \int_0^\infty 2e^{-2x}\,dx = -\Big[e^{-2x}\Big]_0^\infty = 1 \tag{124}
$$

Note that the mean of an exponential with parameter λ is $\frac{1}{\lambda}$ and the variance is $\frac{1}{\lambda^2}$. These results can be established by noting that $M^{(1)}(0)=\mathbb{E}[X]$ and $M^{(2)}(0)=\mathbb{E}[X^2]$ and $\text{var}[X]=\mathbb{E}[X^2]-\Big(\mathbb{E}[X]\Big)^2$.

Suppose $a=0.3$ and do a simulation to verify that you do get an exponential distribution.

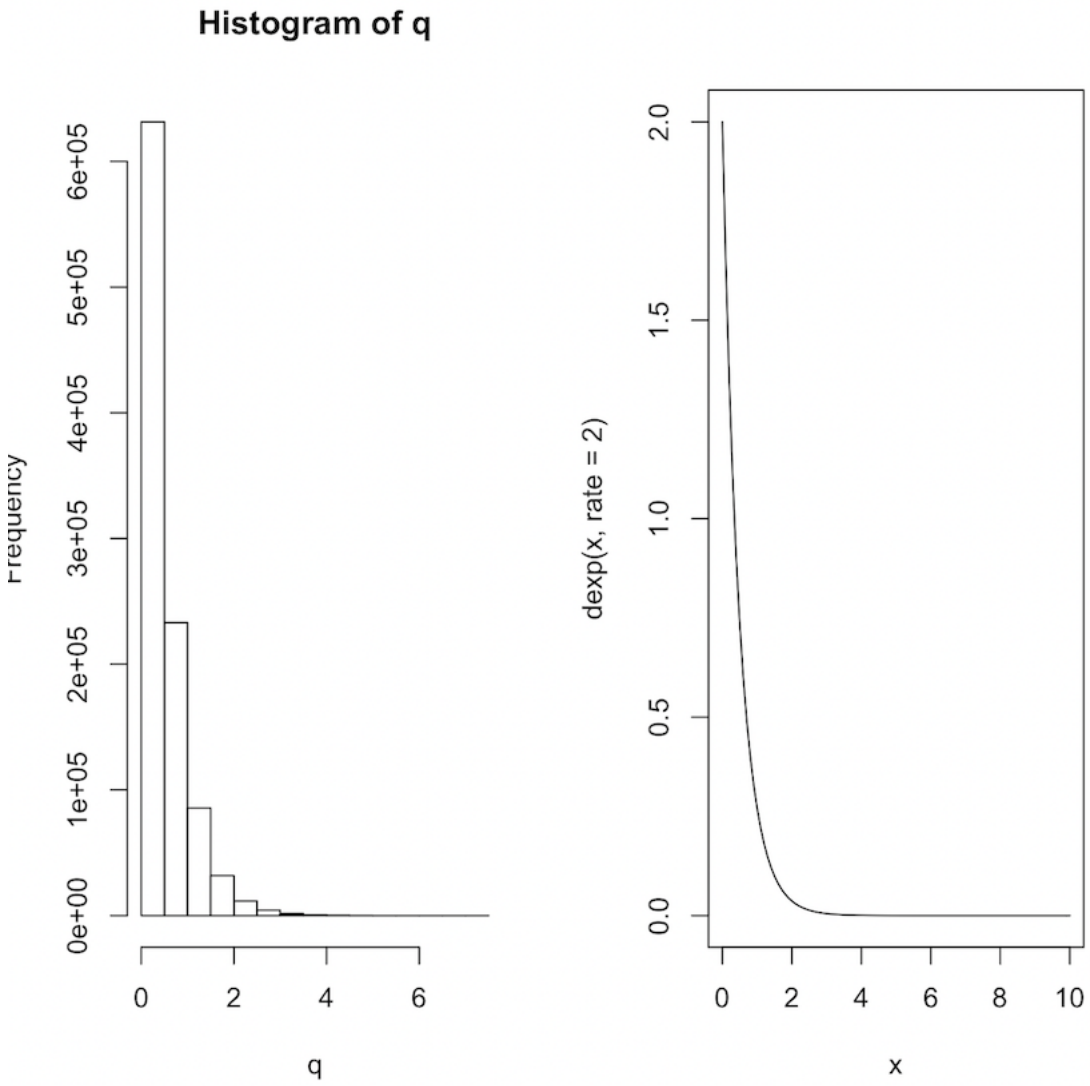
A simulation in Mathematica shows that when we generate 10^6 uniform variates on $(0.3,1)$ and then plug them into $Q=-\frac{1}{2}\ln\left(\frac{X-a}{1-a}\right)$. we get a mean of $\frac{1}{2}$ and a variance of $\frac{1}{4}$ in line with an exponential distribution with $\lambda=2$.

```
In[5]:= unf = Table[RandomVariate[UniformDistribution[{0.3, 1}]], 10^6];

In[6]:= q = -0.5 * Log[ (unf - 0.3)/0.7 ];

In[7]:= Mean[q]

Out[7]= 0.500668

In[8]:= Variance[q]

Out[8]= 0.250194
```

In R the simulation is similar:

```
> v <- runif(10^6,0.3,1)
> q <- -0.5* log( (v-0.3)/0.7)
> mean(q)
[1] 0.4996236
> var(q)
[1] 0.249953
>
```

The histogram of the simulated variates and the theoretical density $f(x) = 2e^{-2x}$ are shown below:



The code in R for the graphs is as follows:

```
par(mfrow=c(1,2))
v <- runif(10^6,0.3,1)
q <- -0.5 * log( (v-0.3)/0.7  )
hist(q)
curve(dexp(x,rate=2),from=0,to=10)
```

Problem 2

Let X_1 and X_2 be two independent standard normal random variates. Let $Y_1 = g_1(X_1, X_2) = X_1 + X_2$ and $Y_2 = g_2(X_1, X_2) = X_2 - X_1$. Find the joint distribution of Y_1 and Y_2.

Solution

This is the type of problem where we should be able to manipulate expectations to get a useful result.

$$\begin{aligned}
M_{Y_1,Y_2}(s_1, s_2) =& \mathbb{E}[e^{Y_1 s_1 + Y_2 s_2}] \\
=& \mathbb{E}[e^{(X_1+X_2)s_1 + (X_2 - X_1)s_2}] \\
=& \mathbb{E}[e^{X_1(s_1 - s_2) + X_2(s_1 + s_2)}] \\
=& \mathbb{E}[e^{X_1(s_1 - s_2)}]\,\mathbb{E}[e^{X_2(s_1 + s_2)}] \\
=& M_{X_1}(s_1 - s_2) M_{X_2}(s_1 + s_2) \\
=& e^{\frac{(s_1 - s_2)^2}{2}} e^{\frac{(s_1 + s_2)^2}{2}} \\
=& e^{s_1^2 + s_2^2} \\
=& e^{\frac{2s_1^2}{2}} e^{\frac{2s_2^2}{2}} \\
=& M_{Y_1}(s_1) M_{Y_2}(s_2)
\end{aligned} \tag{125}$$

Note that $M_{aU+bV}(s) = M_U(as) \centerdot M_V(bs)$ where U, V are independent. Hence $M_{X_1+X_2}(s) = M_{X_1}(s) \centerdot M_{X_2}(s)$ so $e^{\frac{2s_1^2}{2}} = M_{X_1}(s_1) \centerdot M_{X_2}(s_1) = M_{Y_1}(s_1)$ since X_1 and X_2 are $N(0,1)$. Note that you get the same result for $Y_2 = X_2 - X_1$ because $M_{X_1}(-s) = M_{X_1}(s)$. To prove this just integrate $\int_{-\infty}^{\infty} e^{-sx} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx$ as this represents $M_{X_1}(-s)$

Problem 3

Let X_1 and X_2 be two independent standard normal random variables. Let $Y = \frac{(X_1 - X_2)^2}{2}$ and hence find the distribution of Y.

Solution

In order to solve this problem you need to use (116) and the fact that the density for a $N(0,1)$ variable is $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ and that $\int_{\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx = 1$. Furthermore the independence of the variables implies that the expectation of a product is the product of the expectations.

$$
\begin{aligned}
M_Y(s) =& \mathbb{E}[e^{Ys}] \\
=& \mathbb{E}[e^{\frac{(X_1-X_2)^2}{2}s}] \\
=& \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\frac{(x_1-x_2)^2}{2}s}\frac{1}{\sqrt{2\pi}}e^{-\frac{x_1^2}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x_2^2}{2}}\,dx_1dx_2 \\
=& \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi}e^{\left[\frac{(x_1-x_2)^2 s}{2}-\frac{(x_1^2+x_2^2)}{2}\right]}\,dx_1dx_2 \\
=& \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi}e^{-\frac{1}{2}\left[x_1^2(1-s)+2x_1x_2s+x_2^2(1-s)\right]}\,dx_1dx_2 \\
=& \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x_2^2(1-s)}\times\left\{\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\left[\frac{(1-s)}{2}\left[x_1^2+\frac{2x_1x_2s}{1-s}\right]\right]}\,dx_1\right\}dx_2 \\
=& \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x_2^2(1-s)}\times\left\{\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\left[\frac{(1-s)}{2}\left[\left(x_1+\frac{x_2s}{1-s}\right)^2-\frac{x_2^2s^2}{(1-s)^2}\right]\right]}\,dx_1\right\}dx_2 \\
=& \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x_2^2(1-s)}\times e^{\frac{x_2^2s^2}{2(1-s)}}\times\left\{\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\left[\frac{(1-s)}{2}\left(x_1+\frac{x_2s}{1-s}\right)^2\right]}\,dx_1\right\}dx_2 \\
=& \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x_2^2(1-s)}\times e^{\frac{x_2^2s^2}{2(1-s)}}\times\left\{\frac{1}{\sqrt{1-s}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}\,du\right\}dx_2 \quad u=\sqrt{1-s}\left(x_1+\frac{x_2s}{1-s}\right) \\
=& \frac{1}{\sqrt{1-s}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x_2^2(1-s)}\,e^{\frac{x_2^2s^2}{2(1-s)}}\,dx_2 \\
=& \frac{1}{\sqrt{1-s}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x_2^2}{2(1-s)}\left[(1-s)^2-s^2\right]}\,dx_2 \\
=& \frac{1}{\sqrt{1-s}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x_2^2}{2(1-s)}(1-2s)}\,dx_2 \quad u=\sqrt{\frac{1-2s}{1-s}}x_2 \\
=& \frac{1}{\sqrt{1-s}}\sqrt{\frac{1-s}{1-2s}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}\,du \\
=& \frac{1}{\sqrt{1-2s}} \\
=& \left(\frac{\frac{1}{2}}{\frac{1}{2}-s}\right)^{\frac{1}{2}} \text{ for } s<\frac{1}{2}
\end{aligned}
\tag{126}
$$

Recall that the Gamma distribution $f_X(x,r,\lambda) = \frac{\lambda}{\Gamma(r)}(\lambda x)^{r-1}e^{-\lambda x}\mathbb{I}_{(0,\infty)}(x)$. Thus (126) is recognizable as a Gamma function with parameters $r=\frac{1}{2}$ and $\lambda=\frac{1}{2}$ so that $f_Y(y) = \frac{\sqrt{\frac{1}{2}}}{\Gamma(\frac{1}{2})}y^{-\frac{1}{2}}e^{-\frac{y}{2}}\mathbb{I}_{(0,\infty)}(y)$. Note that $\mathbb{I}_A(x)=1$ if $x\in A$ and 0 otherwise.

Problem 4

(a) Assuming X and Y are continuous random variables derive the distribution of the sum $Z=X+Y$ and its density. What is the probability density function (pdf) of Z when X and Y

are independent?

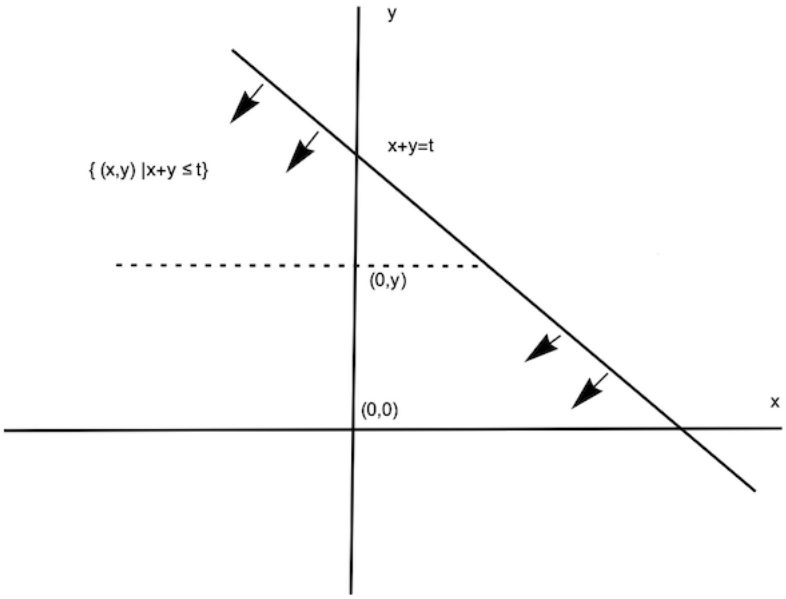
(b) Suppose that Y_1 is uniform on $[0, a]$ and Y_2 is uniform on $[0, a]$ for $a > 0$ and that Y_1 and Y_2 are independent. Is $Y_1 + Y_2$ uniform on $[0, 2a]$ and, if not, why not? Show your reasoning both analytically and running a simulation with $a = 3$.

Solution

(a) The distribution of $Z = X + Y$ is:

$$F_Z(t) = F_{X+Y}(t) = \mathbb{P}[X + Y \leq t] \tag{127}$$

The following figure sets out the region of integration:



The required probabiity $\mathbb{P}[X + Y \leq t]$ is the probability that (x, y) falls within the region below the line $x + y = t$ in the figure. Thus the probability is just the integral of the density over the region ie:

$$F_{X+Y}(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{t-y} f(x, y)\, dx\, dy \tag{128}$$

where $\infty < t < \infty$.

By making the subtitution $u = x + y$ where y is fixed, we have that $dx = du$ and if $x = t - y$ we have that $u = t$. Hence:

$$\int_{-\infty}^{t-y} f(x,y)\,dx = \int_{-\infty}^{t} f(u-y,y)\,du \tag{129}$$

and so:

$$\begin{aligned} F_{X+Y}(t) &= \int_{-\infty}^{\infty} \Big(\int_{-\infty}^{t} f(u-y,y)\,du \Big)\,dy \\ &= \int_{-\infty}^{t} \Big(\int_{-\infty}^{\infty} f(u-y,y)\,dy \Big) du \end{aligned} \tag{130}$$

So the distribution function we are after is:

$$\boxed{F_{X+Y}(t) = \int_{-\infty}^{t} \Big(\int_{-\infty}^{\infty} f(u-y,y)\,dy \Big) du} \tag{131}$$

We differentiate (131) to get the pdf $f_{X+Y}(t)$:

$$\boxed{f_{X+Y}(t) = \int_{-\infty}^{\infty} f(t-y,y)\,dy \text{ for } -\infty < t < \infty} \tag{132}$$

Now if X and Y are independent we have that:

$$\boxed{f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(t-y) f_Y(y)\,dy \text{ for } -\infty < t < \infty} \tag{133}$$

which is the **convolution** of f_X and f_Y.

(b) Let $Z = X + Y$. Using indicator functions the densities of X and Y are as follows:

$$f_X(x) = \frac{1}{a}\mathbb{I}_{[0,a]}(x) \tag{134}$$

and

$$f_Y(y) = \frac{1}{a}\mathbb{I}_{[0,a]}(y) \tag{135}$$

where $\mathbb{I}_A(x) = 1$ if $x \in A$ and 0 for $x \notin A$

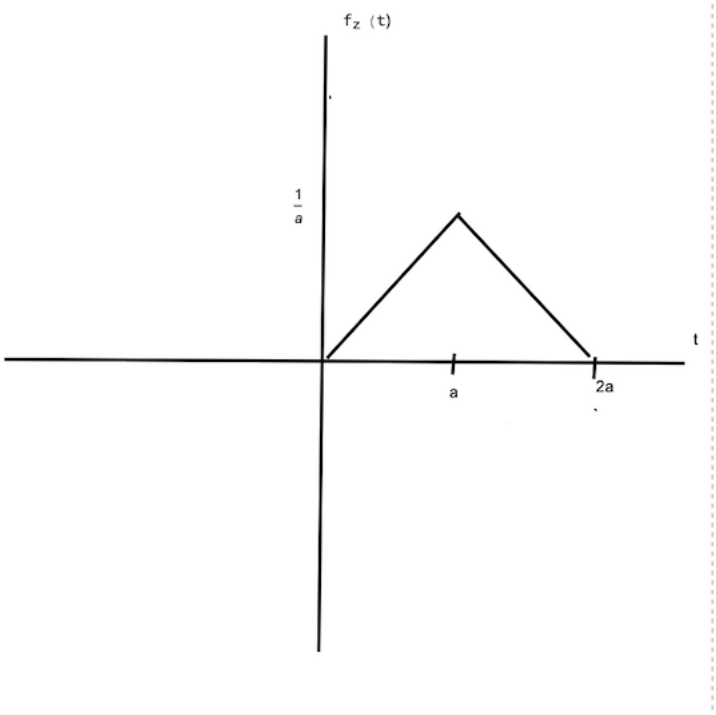
Using (133) we have that $f_Z(t) = f_{X+Y}(t)$ is:

$$\begin{aligned} f_{X+Y}(t) &= \int_{-\infty}^{\infty} f_X(t-y) f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} \frac{1}{a}\mathbb{I}_{[0,a]}(t-y)\frac{1}{a}\mathbb{I}_{[0,a]}(y)\, dy \\ &= \frac{1}{a^2}\int_{-\infty}^{\infty} \mathbb{I}_{[0,a]}(t-y)\mathbb{I}_{[0,a]}(y)\, dy \end{aligned} \tag{136}$$

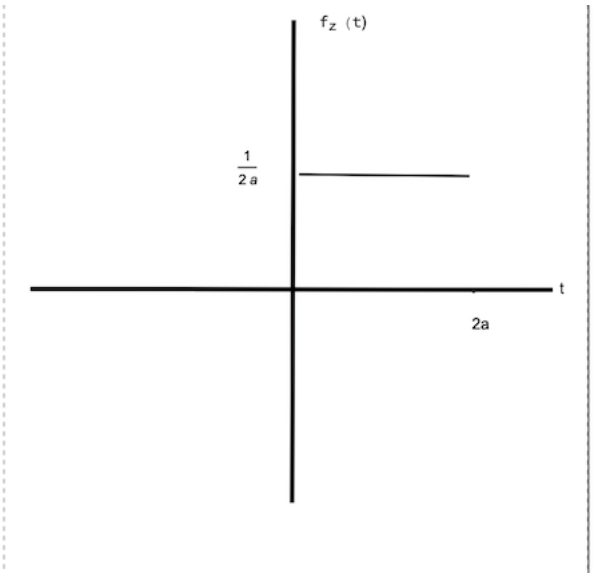
Now noting that $t \in [0, 2a]$ we break the domain of integration into two intervals $[0, a]$ and $[a, 2a]$ as follows (a pedantic point would be that the intervals are technically not disjoint because a is in both sets but for integration purposes the integral is not affected because this single point is of measure zero):

$$\begin{aligned} f_{X+Y}(t) =& \frac{1}{a^2}\int_{-\infty}^{\infty} \mathbb{I}_{[0,a]}(t-y)\mathbb{I}_{[0,a]}(y)\, dy \\ =& \frac{1}{a^2}\int_{-\infty}^{\infty} \left\{\mathbb{I}_{[0,t]}(y)\mathbb{I}_{[0,a]}(t) + \mathbb{I}_{[t-a,a]}(y)\mathbb{I}_{[a,2a]}(t)\right\} dy \\ =& \frac{1}{a^2}\mathbb{I}_{[0,a]}(t)\int_0^t dy + \frac{1}{a^2}\mathbb{I}_{[a,2a]}(t)\int_{t-a}^{a} dy \\ =& \frac{t}{a^2}\mathbb{I}_{[0,a]}(t) + \frac{2a-t}{a^2}\mathbb{I}_{[a,2a]}(t) \end{aligned} \tag{137}$$

The graph of (137) is:

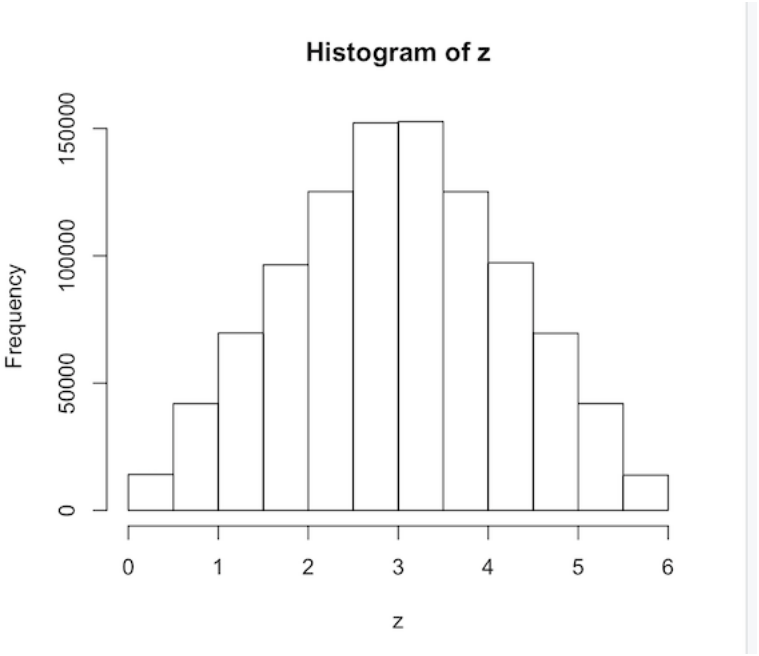


Thus Z is not uniform on $[0, 2a]$, the graph of whose density would look like this:



We can run a simple simulation in RStudio to find the mean and variance of Z and also generate a histogram. This is what we get:

```
> u<-runif(10^6,0,3)
> v<- runif(10^6,0,3)
> z<-u+v
> hist(z)
> mean(z)
[1] 3.00041
> var(z)
[1] 1.504249
```



If Z were uniform on $[0, 6]$ its mean would be:

$$\mathbb{E}[Z] = \frac{6-0}{2} = 3 \tag{138}$$

and its variance would be:

$$\text{var}[Z] = \frac{(6-0)^2}{12} = 3 \tag{139}$$

The simulated variance is 1.5.

We can analytically derive the mean and variance of Z using (137):

$$\begin{aligned}\mathbb{E}[Z] =& \int_0^6 t f_Z(t)\,dt \\ =& \frac{1}{9}\int_0^3 t^2\,dt + \frac{1}{9}\int_3^6 t(6-t)\,dt \\ =& 3\end{aligned} \tag{140}$$

$$\begin{aligned}\mathbb{E}[Z^2] =& \int_0^6 t^2 f_Z(t)\,dt \\ =& \frac{1}{9}\int_0^3 t^3\,dt + \frac{1}{9}\int_3^6 t^2(6-t)\,dt \\ =& \frac{21}{2}\end{aligned} \tag{141}$$

Hence $\text{var}[Z] = \mathbb{E}[Z^2] - \Big(\mathbb{E}[Z]\Big)^2 = \frac{21}{2} - 3^2 = 1.5$. Which is what the simulation gave.

Problem 5

Let X be a random variable which has mean $\mu \neq 0$ and variance σ_1^2. Let Y be another random variable which has the same mean μ and variance σ_2^2 and further suppose that the correlation between X and Y is ρ where $-1 \leq \rho \leq 1$. Let $Q = aX + bY$ where a and b are any real numbers and suppose it is known that $\mathbb{E}[Q] = \mu$. Determine the values of a and b so that the variance of Q is minimised.

Solution

We have:

$$\begin{aligned}\mathbb{E}[X] =& \mu \neq 0 \\ \text{var}[X] =& \sigma_1^2 \\ \mathbb{E}[Y] =& \mu \neq 0 \\ \text{var}[Y] =& \sigma_2^2 \\ \rho =& \frac{\text{cov}(X,Y)}{\sigma_1\sigma_2}\end{aligned} \tag{142}$$

From $\mathbb{E}[Q] = \mu$ we have:

$$\begin{aligned}\mathbb{E}[Q] =&\mathbb{E}[aX + bY] \\ =&a\mathbb{E}[X] + b\mathbb{E}[Y] \text{ using linearity of } \mathbb{E} \\ =&a\mu + b\mu = \mu \\ \implies& a + b = 1 \text{ since } \mu \neq 0\end{aligned} \tag{143}$$

$$\begin{aligned}\text{var}[Q] =&\text{var}[aX + bY] \\ =&a^2\text{var}[X] + b^2\text{var}[Y] + 2ab\,\text{cov}(X, Y) \\ =&a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\sigma_1\sigma_2\rho \\ =&a^2\sigma_1^2 + (1-a)^2\sigma_2^2 + 2a(1-a)\sigma_1\sigma_2\rho \text{ using (143)}\end{aligned} \tag{144}$$

To find a turning point we want $\frac{d\text{var}[Q]}{da} = 0$:

$$\frac{d\text{var}[Q]}{da} = 2a\sigma_1^2 - 2(1-a)\sigma_2^2 + (2-4a)\sigma_1\sigma_2\rho = 0 \tag{145}$$

Therefore:

$$\begin{aligned}a(2\sigma_1^2 + 2\sigma_2^2 - 4\sigma_1\sigma_2\rho) =&2\sigma_2^2 - 2\sigma_1\sigma_2\rho \\ a =&\frac{2\sigma_2^2 - 2\sigma_1\sigma_2\rho}{2\sigma_1^2 + 2\sigma_2^2 - 4\sigma_1\sigma_2\rho} \\ a =&\frac{\sigma_2^2 - \sigma_1\sigma_2\rho}{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho}\end{aligned} \tag{146}$$

and so:

$$\begin{aligned}b =&1 - a \\ =&1 - \frac{\sigma_2^2 - \sigma_1\sigma_2\rho}{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho} \\ =&\frac{\sigma_1^2 - \sigma_1\sigma_2\rho}{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho}\end{aligned} \tag{147}$$

For a minimum $\frac{d^2\text{var}[Q]}{da^2} \geq 0$, so checking for this:

$$\begin{aligned}\frac{d^2\text{var}[Q]}{da^2} =& 2\sigma_1^2 + 2\sigma_2^2 - 4\sigma_1\sigma_2\rho \\ =& 2(\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho) \\ \geq& \sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2 \text{ since } \rho \leq 1 \implies -2\sigma_1\sigma_2\rho \geq -2\sigma_1\sigma_2 \\ =& (\sigma_1 - \sigma_2)^2 \\ \geq& 0\end{aligned} \tag{148}$$

Problem 6

There are two testing laboratories A and B which test for a certain condition. Lab A has a probability of 0.98 of correctly identifying the condition and Lab B has a probability of 0.985 of correctly identifying the condition. If samples are given to both labs for independent testing, what is the probability of the condition being correctly identified? Show your reasoning.

Solution
Let A = the event that Lab A correctly identifies the condition and B= the event that Lab B correctly identifies the condition. The complements of these events are $\bar{A}$ and $\bar{B}$.

For intersections of sets we write AB etc and for unions we write $A + B$.

The required probability is as follows, using the independence of the testing:

$$\begin{aligned}\mathbb{P}[AB + A\bar{B} + \bar{A}B] =& \mathbb{P}[AB] + \mathbb{P}[A\bar{B}] + \mathbb{P}[\bar{A}B] \\ =& 0.98 \times 0.985 + 0.98 \times 0.015 + 0.02 \times 0.985 \\ =& 0.9997\end{aligned} \tag{149}$$

Problem 7

Suppose Y_1 is Poisson (λ_1) and Y_2 is Poisson (λ_2) where Y_1 and Y_2 are independent, then $Y_1 + Y_2$ is Poisson $(\lambda_1 + \lambda_2)$. True or false? Give your reasons.

Solution

Independent Poisson processes reproduce under summation so the statement is true. The moment generating functions for Y_1 and Y_2 are:

$$\begin{aligned}m_{Y_1}(t) =& e^{\lambda_1(e^t - 1)} \\ m_{Y_2}(t) =& e^{\lambda_2(e^t - 1)}\end{aligned} \tag{150}$$

If $Y = Y_1 + Y_2$ then the mgf for Y is (using independence):

$$\begin{aligned} m_Y(t) =& e^{\lambda_1(e^t-1)} \times e^{\lambda_2(e^t-1)} \\ =& e^{(\lambda_1+\lambda_2)(e^t-1)} \end{aligned} \tag{151}$$

ie this is Poisson $(\lambda_1 + \lambda_2)$.

Problem 8

Let $X_1, X_2, \ldots, X_n$ be mutually independent random variables such that each X_k assumes the values 1 and 0 with probabilities p_k and $q_k = 1-p_k$ respectively. Let $S_n = X_1+X_2+\cdots+X_n$. Show that:

(1) $\text{var}(S_n) = \sum_{k=1}^n p_k q_k$

(2) S_n can be interpreted as the total number of successes in n independent trials, each of which results in success or failure. Then $p = \frac{p_1+\cdots+p_n}{n}$ is the average probability of success. Now consider a Bernoulli process with the constant probability of success p. Show that $\text{var}(S_n) = np - \sum_{k=1}^n p_k^2$.

(3) Show by either induction or Lagrange multipliers that among all combinations $\{p_k\}$ such that $\sum_{k=1}^n p_k = np$ the sum $\sum_{k=1}^n p_k^2$ assumes its minimum value when all the p_k are equal. Conclude that, if the average probability of success p is kept constant, $\text{var}(S_n)$ assumes its maximum when $p_1 = p_2 = \cdots = p_n = p$.

(4) Comment on this observation:"The variability of p_k, or lack of uniformity, decreases the magnitude of chance fluctuations as measured by the variance" . Is this surprising?

Solution

(1) We have to show that $\text{var}(S_n) = \sum_{k=1}^n p_k q_k$.

For each k we have that the expectation of X_k, $\mathbb{E}[X_k] = p_k$ and the variance:

$$\text{Var}(X_k) = \mathbb{E}[X_k^2] - (\mathbb{E}[X_k])^2 = p_k - p_k^2 = p_k q_k \tag{152}$$

Note that $\mathbb{E}[X_k^2] = 1^2 \times p_k + 0^2 \times q_k = p_k$.

We know that the variance of the sum of n mutually independent random variables X_k, $S_n = \sum_{k=1}^n X_k$ is:

$$\text{Var}(S_n) = \sum_{k=1}^n \sigma_k^2 \quad \text{where: } \sigma_k = \sqrt{\text{Var}(X_k)} \tag{153}$$

Hence, using (152) we have:

$$\mathrm{Var}(S_n) = \sum_{k=1}^{n} \mathrm{Var}(X_k) = \sum_{k=1}^{n} p_k q_k \tag{154}$$

(2) We have to show that: $\mathrm{var}(S_n) = np - \sum_{k=1}^{n} p_k^2$ where $p = \frac{p_1 + \cdots + p_n}{n}$

This is simply a rewriting of (154) as follows:

$$\begin{aligned} \mathrm{Var}(S_n) =& \sum_{k=1}^{n} p_k q_k \\ =& \sum_{k=1}^{n} p_k (1 - p_k) \\ =& \sum_{k=1}^{n} p_k - \sum_{k=1}^{n} p_k^2 \\ =& np - \sum_{k=1}^{n} p_k^2 \end{aligned} \tag{155}$$

(3)

Using Lagrange multipliers we have the folllowing. We minimise $\sum_{k=1}^{n} p_k^2$ in order to maximise the variance:

Minimise $\sum_{k=1}^{n} p_k^2$ subject to the constraint: $p = \frac{p_1 + \cdots + p_n}{n}$ is constant ie $\sum_{k=1}^{n} p_k = np$

In the language of Lagrange multipliers we thus have something of the form of minimising $F(p_1, p_2, \ldots, p_n)$ subject to $\phi(p_1, p_2, \ldots, p_n) = 0$. Here $F(p_1, p_2, \ldots, p_n) = \sum_{k=1}^{n} p_k^2$ and $\phi(p_1, p_2, \ldots, p_n) = \sum_{k=1}^{n} p_k - np = 0$. Thus we form the auxiliary function:

$$G(p_1, p_2, \ldots, p_n) = F(p_1, p_2, \ldots, p_n) + \lambda \phi(p_1, p_2, \ldots, p_n) \tag{156}$$

To find an extremum we need:

$$\frac{\partial G}{\partial p_k} = 0 \quad \forall k \tag{157}$$

Recall that this is a necessary condition so further investigation is needed to establish that we actually have a minimum.

Making the relevant substitutions in (157) we have:

$$G(p_1, p_2, \ldots, p_n) = \sum_{k=1}^{n} p_k^2 + \lambda\big(\sum_{k=1}^{n} p_k - np\big) \tag{158}$$

Differentiating we get:

$$\frac{\partial G}{\partial p_k} = 2p_k + \lambda = 0 \implies \lambda = -2p_k \quad \forall k \tag{159}$$

Now (159) only makes sense when the p_k are constant ie $p_k = p^*$ for all k. Recall from the theory of Lagrange multipliers in 2 dimensions that you have something like $\nabla f(x,y) = \lambda \nabla g(x,y)$ so that you need a constant λ (ignoring sign) for the normal to the surface. The constraint $\sum_{k=1}^{n} p_k - np = 0$ then becomes $\sum_{k=1}^{n} p^* - np = 0$. Therefore $np^* - np = 0$ and so $p^* = p = \frac{p_1 + \cdots + p_n}{n}$.

To see that $p = \frac{p_1 + \cdots + p_n}{n} = p_k \ \forall k$ actually minimises $\sum_{k=1}^{n} p_k^2$ perturb two of the p_k as follows. Without loss of generality we can relabel the p_k so that $p_1 \leq p_2 \leq \cdots \leq p_n$. Let:

$$p_1^* = p_1 - \epsilon, \quad p_2^* = p_2 + \epsilon \tag{160}$$

where $\epsilon > 0$. The other values of p_k remain the same ie $p_k^* = p_k$ for $k \neq 1, 2$. Hence $np = np^* = n\sum_{k=1}^{n} p_k^*$ ie it is constant. Thus:

$$\sum_{k=1}^{n} p_k^2 - \sum_{k=1}^{n} {p_k^*}^2 = \sum_{k=1}^{n} p_k^2 - \big((p_1 - \epsilon)^2 + (p_2 + \epsilon)^2) + p_3^2 + \cdots + p_n^2\big) = -2\epsilon^2 - 2\epsilon(p_2 - p_1) < 0 \tag{161}$$

Noting the assumption that $p_1 \leq p_2$. Thus $\sum_{k=1}^{n} p_k^2 < \sum_{k=1}^{n} {p_k^*}^2$ and so we do indeed have a minimum. A more rigorous and general approach to proving the minimum involves Hessians in n dimensions.

To prove the property by induction we have to show that $p = \frac{p_1 + \cdots + p_n}{n}$ minimises $\sum_{k=1}^{n} p_k^2$. The base case of $n = 1$ is trivial (and bereft of useful insight) so we consider $n = 2$. Thus:

$$\sum_{k=1}^{2} p_k^2 - \sum_{k=1}^{2} p^2 = p_1^2 + p_2^2 - 2\Big(\frac{p_1 + p_2}{2}\Big)^2 = \frac{p_1^2 - 2p_1p_2 + p_2^2}{2} = \frac{(p_1 - p_2)^2}{2} \geq 0 \tag{162}$$

Thus $\sum_{k=1}^{2} p^2$ is minimal. Now assuming the property is true for all n, we consider the situation for $n + 1$ where $p^* = \frac{p_1 + \cdots + p_n + p_{n+1}}{n+1}$:

$$\begin{aligned}
\sum_{k=1}^{n+1} p_k^2 - \sum_{k=1}^{n+1} p^{*\,2} &= \sum_{k=1}^{n+1} p_k^2 - (n+1)\frac{(p_1 + \cdots + p_n + p_{n+1})^2}{(n+1)^2} \\
&= \sum_{k=1}^{n} p_k^2 + p_{n+1}^2 - \frac{(np + p_{n+1})^2}{n+1} \\
&\geq np^2 + p_{n+1}^2 - \frac{(n^2p^2 + 2n\,p\,p_{n+1} + p_{n+1}^2)}{n+1} \\
&\quad \text{using the induction hypothesis that } \sum_{k=1}^{n} p_k^2 \geq \sum_{k=1}^{n} p^2 \text{ and } np = \sum_{k=1}^{n} p_k \text{ is constant} \\
&= \frac{n(n+1)p^2 + (n+1)p_{n+1}^2 - n^2p^2 - 2n\,p\,p_{n+1} - p_{n+1}^2}{n+1} \\
&= \frac{n^2p^2 + np^2 + np_{n+1}^2 + p_{n+1}^2 - n^2p^2 - 2n\,p\,p_{n+1} - p_{n+1}^2}{n+1} \\
&= \frac{np^2 - 2n\,p\,p_{n+1} + np_{n+1}^2}{n+1} \\
&= \frac{n(p - p_{n+1})^2}{n+1} \geq 0
\end{aligned} \tag{163}$$

Thus $\sum_{k=1}^{n+1} p_k^2 \geq \sum_{k=1}^{n+1} p^{*\,2}$ and so the proposition is established by induction.

(4)

The quote comes from Feller ([3], page 231.) In their book **"One Thousand Exercises in Probability"**, Oxford University Press, 2003, Geoffrey Grimmett and David Stirzaker cheekily say at page 176 that "This conclusion is not contrary to informed intuition, but experience shows is to be contrary to much uninformed intuition". Their proof without Lagrange multipliers is as follows.

Let $s = \sum_{k=1}^{n} p_k$ and let Z be a random variable taking each of the values of $p_1, p_2, \ldots, p_n$ with equal probability $\frac{1}{n}$. Now $\mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 = \mathrm{var}(Z) \geq 0$ so that:

$$\sum_{k=1}^{n} \frac{1}{n} p_k^2 \geq \Big(\sum_{k=1}^{n} \frac{1}{n} p_k \Big)^2 = \frac{s^2}{n^2} \tag{164}$$

with equality if and only if Z is (almost surely) constant, which is to say that $p_1 = p_2 = \cdots = p_n$. Hence:

$$\mathrm{var}(Y) = \sum_{k=1}^{n} p_k - \sum_{k=1}^{n} p_k^2 \leq s - \frac{s^2}{n} \tag{165}$$

with equality if and only if $p_1 = p_2 = \cdots = p_n$.

The reason this problem seems counter-intuitive at first blush is that we may perceive a "big" variance as being associated with big variations in the relevant probabilities. But in the Bernoulli process the mean is simply p_k ie $\mathbb{E}[X_k] = p_k$ and we are looking at the expectation of the squared deviation of this with the scores 0 and 1. When you do the calculation you get that the variance of the sum is $\sum_{k=1}^{n} p_k - \sum_{k=1}^{n} p_k^2 = \sum_{k=1}^{n} p_k q_k$. Thus the variance is the simply the sum of the product of the probabilities and their "conjugates" ie $q_k = 1 - p_k$. If you think of the p_k as vectors which are independent and hence orthogonal, to maximise the area spanned by the vectors you make them the same length (the cross product of the two orthogonal vectors would simply be the product of the two lengths which is an area and is maximal when you get a square ie both lengths are the same.

To convince yourself this problem is not hocus pocus, consider the case of $n = 2$ and let $p_1 = 0.1$ and $p_2 = 0.9$ so that there is a big difference in probabilities. Thus we have $p = \frac{p_1+p_2}{2} = 0.5$ and the variance is $1 - (0.1^2 + 0.9^2) = 0.18$. If $p_1 = p_2 = p$ then the variance is $2p - 2p^2 = 2p(1-p) = 0.5$ which exceeds 0.18.

Now choose $p_1 = 0.49$ and $p_2 = 0.51$ so that the sum of the probabilities is the same as the first case, but the probabilities are much closer. Again $p = 0.5$ but the variance is $1 - (0.49^2 + 0.51^2) = 0.4998$. Calculating the variance with $p = 0.5$ gives 0.5 as before which is greater than 0.4998

See [6] for a discussion of how this result applies in the context of maximising Shannon entropy.

Problem 9

(a) Let $f(.)$ be some density with mean μ and finite variance σ and let $\bar{X}_n$ be the sample mean of a random sample of size n taken from $f(.)$. Let ϵ and δ be any two specified numbers satisfying $\epsilon > 0$ and $0 < \delta < 1$. The Weak Law of Large Numbers (WLLN) states that:

$$\mathbb{P}[-\epsilon < \bar{X}_n - \mu < \epsilon] \geq 1 - \delta \tag{166}$$

(a) You know that the WLLN can be used to work out sample sizes and you remember equation (166) above. You also recall that the proof of the WLLN uses Chebyshev's inequality which you do remember in the following form:

$\mathbb{P}[g(X) \geq k] \leq \frac{\mathbb{E}[g(x)]}{k}$ for every $k > 0$, every random variable X and non-negative function $g(.)$.

Using Chebyshev's inequality prove the WLLN by using $g(X) = (\bar{X}_n - \mu)^2$ and $k = \epsilon^2$

(b) Now suppose that some distribution with an unknown mean has a variance equal to 1. How large a sample must be taken in order that the probability will be at least 0.95 that the sample mean $\bar{X}_n$ will lie within 0.5 of the population mean? Show your working.

Solution

(a) We let $g(X) = (\bar{X}_n - \mu)^2$ and $k = \epsilon^2$. Then:

$$\begin{aligned}
\mathbb{P}[-\epsilon < \bar{X}_n - \mu < \epsilon] =& \mathbb{P}[|\bar{X}_n - \mu| < \epsilon] \\
=& \mathbb{P}[|\bar{X}_n - \mu|^2 < \epsilon^2] \\
\geq& 1 - \frac{\mathbb{E}[(\bar{X}_n - \mu)^2]}{\epsilon^2} \\
=& 1 - \frac{\left(\frac{1}{n}\right)\sigma^2}{\epsilon^2} \\
\geq& 1 - \delta
\end{aligned} \tag{167}$$

This will hold for $\delta > \frac{\sigma^2}{n\epsilon^2}$ or $n > \frac{\sigma^2}{\epsilon^2 \delta}$

Note that $\mathbb{P}[Z \leq z] = 1 - \mathbb{P}[Z < z]$

The fundamental step in this analysis is the fact that $\text{var}[\bar{X}_n] = \frac{\sigma^2}{n}$. Note that the variance of the sum of n independent, identically distributed variables X_i with mean μ_X and variance σ_X and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ then:

$$\mathbb{E}[\bar{X}_n] = \mu_X \tag{168}$$

and

$$\text{var}[\bar{X}_n] = \frac{\sigma_X^2}{n} \tag{169}$$

Recall that $\text{var}\left[\sum_{j=1}^{n} a_j X_j\right] = \sum_{j=1}^{n}\sum_{k=1}^{n} a_j\, a_k \,\text{cov}[X_j, X_k] = \sum_{j=1}^{n} a_j^2 \,\text{var}[X_j] + \sum\sum_{j \neq k} \text{cov}[X_j, X_k]$. Because the variables are independent the covariance is zero and because they are identically distributed we get $\frac{n}{n^2}\sigma_X^2 = \frac{1}{n}\sigma_X^2$

Note that although the independent, identically distributed criteria were not explicitly stated they are implicit in the concept of a random sample of the population at issue.

(b) The hypotheses of the WLLN are satisfied and we can apply it using $\sigma^2 = 1$ and $\epsilon = 0.5$ and $\delta = 0.05$ so that we need a sample n such that:

$$n > \frac{\sigma^2}{\epsilon^2\, \delta} = \frac{1}{(0.5)^2 0.05} = 80 \tag{170}$$

Problem 10

(a) Suppose that X_0, X_1 and X_2 are three independent Poisson random variables with distinct parameters. Thus: $X_0 \sim \text{Poi}(\lambda_0)$, $X_1 \sim \text{Poi}(\lambda_1)$ and $X_2 \sim \text{Poi}(\lambda_2)$. Now consider the two variables Y and Z defined as:

$$\begin{aligned} Y =& X_0 + X_1 \\ Z =& X_0 + X_2 \end{aligned} \tag{171}$$

Show that the covariance of Y and Z can never be less than or equal to zero and comment on whether the result surprises you.

(b) A Youtube video on bivariate Poisson distributions sets up the derivation of the bivariate distribution of Y and Z (as given in part(a)) with this statement, where X_0 is another Poisson variable:

$$\mathbb{P}[Y = y, Z = z] = \sum_{i=0}^{\infty} \mathbb{P}[Y = y, Z = z, X_0 = i] \tag{172}$$

Is this correct? Provide a rigorous proof or refutation of this equation.

(c) The complete derivation of the bivariate distribution given in the video is as follows. Is it correct- give reasons?

$$\mathbb{P}[Y = y, Z = z] = \sum_{i=0}^{\infty} \mathbb{P}[Y = y, Z = z, X_0 = i] \tag{173a}$$

$$= \sum_{i=0}^{\infty} \mathbb{P}[Y = y, Z = z | X_0 = i] \, \mathbb{P}[X_0 = i] \tag{173b}$$

$$= \sum_{i=0}^{\infty} \mathbb{P}[X_0 + X_1 = y, X_0 + X_2 = z | X_0 = i] \, \mathbb{P}[X_0 = i] \tag{173c}$$

$$= \sum_{i=0}^{\infty} \mathbb{P}[X_1 = y - i, X_2 = z - i] \, \mathbb{P}[X_0 = i] \tag{173d}$$

$$= \sum_{i=0}^{\min\{y,z\}} \frac{e^{-\lambda_0} \lambda_0^i}{i!} \, \frac{e^{-\lambda_1} \lambda_1^{y-i}}{(y-i)!} \, \frac{e^{-\lambda_2} \lambda_2^{z-i}}{(z-i)!} \tag{173e}$$

$$= e^{-(\lambda_0+\lambda_1+\lambda_2)} \frac{\lambda_1^y}{y!} \, \frac{\lambda_2^z}{z!} \sum_{i=0}^{\min\{y,z\}} \frac{y!}{i!(y-i)!} \, \frac{z!}{i!(z-i)!} i! \left(\frac{\lambda_0}{\lambda_1 \lambda_2} \right)^i \tag{173f}$$

$$= e^{-(\lambda_0+\lambda_1+\lambda_2)} \frac{\lambda_1^y}{y!} \, \frac{\lambda_2^z}{z!} \sum_{i=0}^{\min\{y,z\}} \binom{y}{i} \binom{z}{i} i! \left(\frac{\lambda_0}{\lambda_1 \lambda_2} \right)^i \tag{173g}$$

$$\tag{173h}$$

Solution (a)

We know that $\mathbb{E}[X_0] = \lambda_0$, $\mathbb{E}[X_1] = \lambda_1$ and $\mathbb{E}[X_2] = \lambda_2$. Also because of the inheritance property of Poisson distributions we have that:

$$\begin{aligned}\mathbb{E}[Y] =& \mathbb{E}[X_0 + X_1] \\ =& \mathbb{E}[X_0] + \mathbb{E}[X_1] \\ =& \lambda_0 + \lambda_1\end{aligned} \tag{174}$$

and

$$\begin{aligned}\mathbb{E}[Z] =& \mathbb{E}[X_0 + X_2] \\ =& \mathbb{E}[X_0] + \mathbb{E}[X_2] \\ =& \lambda_0 + \lambda_2\end{aligned} \tag{175}$$

We also know (see (93)) that $\mathbb{E}[X^2]$ for a Poisson variable X is $\lambda^2 + \lambda$

The covariance of Y and Z is:

$$\begin{aligned}\text{cov[Y,Z]} =& \mathbb{E}[YZ] - \mathbb{E}[Y]\,\mathbb{E}[Z] \\ =& \mathbb{E}[(X_0 + X_1)(X_0 + X_2)] - (\lambda_0 + \lambda_1)(\lambda_0 + \lambda_2) \\ =& \mathbb{E}[X_0^2 + X_0X_2 + X_1X_0 + X_1X_2] - (\lambda_0^2 + \lambda_0\lambda_2 + \lambda_1\lambda_0 + \lambda_1\lambda_2) \\ =& \mathbb{E}[X_0^2] + \mathbb{E}[X_0X_2] + \mathbb{E}[X_1X_0] + \mathbb{E}[X_1X_2] - (\lambda_0^2 + \lambda_0\lambda_2 + \lambda_1\lambda_0 + \lambda_1\lambda_2) \\ =& \lambda_0^2 + \lambda_0 + \lambda_0\lambda_2 + \lambda_1\lambda_0 + \lambda_1\lambda_2 - (\lambda_0^2 + \lambda_0\lambda_2 + \lambda_1\lambda_0 + \lambda_1\lambda_2) \\ =& \lambda_0\end{aligned} \tag{176}$$

Because Poisson parameters are strictly positive, ie $\lambda_0 > 0$, it follows that cov[Y,Z] can never be less than or equal to zero. This should come as no surprise because, by definition, Y and Z are correlated since X_0 is a common component and its parameter is always positive.

Solution b

This is correct and it follows from the Theorem of Total Probability. What this theorem (which is essentially a basic set theoretic result) does is to take a set B and look at its intersection with a universe which is the sum of disjoint sets A_i. In other words the A_i are a partition of the sample space (or universe if you were simply doing Boolean set theory). Thus:

$$\cup_{i=1}^{\infty} A_i = \Omega \tag{177}$$

If $\mathbb{P}[A_i] > 0$ for all i then for any event B:

$$\begin{aligned}\mathbb{P}[B] =& \mathbb{P}[A_1 \cap B] + \cdots + \mathbb{P}[A_n \cap B] \\ =& \mathbb{P}[A_1]\,\mathbb{P}[B|A_1] + \cdots + \mathbb{P}[A_n]\,\mathbb{P}[B|A_n]\end{aligned} \tag{178}$$

Now for any event B we have that:

$$\begin{aligned}B =& B \cap \Omega \\ =& B \cap \left(\cup_{i=1}^{\infty} A_i\right) \\ =& \cup_{i=1}^{\infty} \left(B \cap A_i\right)\end{aligned} \tag{179}$$

Note that for a Poisson variable A_i, $\mathbb{P}[A_i] > 0$ for all i. That this is the case flows from the definition of the Poisson distribution: $\frac{e^{-\lambda}\lambda^x}{x!} > 0$ for all $x = 0, 1, 2, \ldots$ and $\lambda > 0$. Thus in our case with each $A_i = X_{0_i}$ the hypotheses of the theorem are satisfied.

Because of disjointess we have that with $B = \{Y = y, Z = z\} = \{Y = y\} \cap \{Z = z\}$ and using (179):

$$\begin{aligned}\mathbb{P}[B] =& \mathbb{P}[\{Y = y\} \cap \{Z = z\}] \\ =& \sum_{i=1}^{\infty} \mathbb{P}[B \cap A_i] \\ =& \sum_{i=1}^{\infty} \mathbb{P}[Y = y, Z = z, X_0 = i]\end{aligned} \tag{180}$$

Also note that:

$$\begin{aligned}\mathbb{P}[B] =& \sum_{i=1}^{\infty} \mathbb{P}[B \cap A_i] \\ =& \sum_{i=1}^{\infty} \mathbb{P}[B|A_i]\,\mathbb{P}[A_i] \\ =& \sum_{i=1}^{\infty} \mathbb{P}[Y = y, Z = z|X_0 = i]\,\mathbb{P}[X_0 = i]\end{aligned} \tag{181}$$

Solution c

Lines 159a and 159b have already been established in part (b). Line 159c is just Y and Z in terms of X_0, X_1 and X_2. Line 159d is just substituting $X_0 = i$ appropriately. Line 159e involves

the observation that X_1 and X_2 can never be negative, hence $y < i$ and $z < i$. This means that i must run from 0 to $\min\{y, z\}$.Lines 159f and 159g are correct manipulations so that the overall result is correct.

8 Appendix

8.1 Proof that independence is functionally inherited

Let us suppose that $X_1, X_2, \ldots X_k$ are independent random variables and that $g_i()$ are k functions such that $Y_j = g_j(X_j)$ for $j = 1, 2, \ldots, k$. Then the $Y_1, Y_2, \ldots, Y_k$ are independent.

Note that if $g_j^{-1}(B_j) = \{z : g_j(z) \in B_j\}$ for $j = 1, 2, \ldots, k$ then the events $\{Y_j \in B_j\}$ and $\{X_j \in g_j^{-1}(B_j)\}$ are equivalent. That this is true is because of basic set theory involving images and pre-images. Hence we have the following:

$$\begin{aligned}\mathbb{P}[Y_1 \in B_1; Y_2 \in B_2; \ldots; Y_k \in B_k] =& \mathbb{P}[X_1 \in g_1^{-1}(B_1); X_2 \in g_2^{-1}(B_2); \ldots, ; X_k \in g_k^{-1}(B_k)] \\ &= \prod_{j=1}^{k} \mathbb{P}[X_j \in g_j^{-1}(B_j)] \\ &= \prod_{j=1}^{k} \mathbb{P}[Y_j \in B_j]\end{aligned} \tag{182}$$

8.2 Sketch of the proof of uniqueness of moment generating functions

The following is based on Chapters 26 and 30 of Billingsley [1]. I fill in all the bits he leaves out but . He starts with the characteristic function defined as follows:

$$\phi(t) = \mathbb{E}[e^{itX}] = \int_{-\infty}^{\infty} e^{itx}\, \mu(dx) \tag{183}$$

where μ is a probability measure. Clearly $\phi(0) = 1$ since the probability integral must equal 1. Also $|\phi(t)| = \left| \int_{-\infty}^{\infty} e^{itx}\, \mu(dx) \right| \leq \int_{-\infty}^{\infty} |e^{itx}|\, \mu(dx) \leq \int_{-\infty}^{\infty} \mu(dx) = 1$. Moreover $\phi(t)$ is continuous in t and because $|\phi(t+h) - \phi(t)| = |\int (e^{i(t+h)x} - e^{itx})\, \mu(dx)| \leq \int |(e^{itx}(e^{ith} - 1)|\, \mu(dx) \leq \int |e^{itx}|\, |e^{ith} - 1|\, \mu(dx) = \int |e^{ith} - 1|\, \mu(dx)$, it follows from the bounded convergence theorem that $\phi(t)$ is uniformly continuous. In the context of Fourier theory none of this is surprising. If a function f is a creature of Schwartz space it is uniformly continuous on any compact interval and so is its Fourier transform (see [5], pages 136-142)

Billingsley gets some bounds on the relevant integral and to do this he uses Taylor's theorem with the integral form of the remainder. The crux of the approach is to get a Taylor's series expansion for e^{ix} with the integral form of the remainder. The representation is (noting that in

the represention the remainder term for a function $f(x)$ is $\frac{1}{n!}\int_0^x (x-t)^n f^{(n+1)}(t)\,dt$):

$$\begin{aligned} e^{ix} =& \sum_{k=0}^{n} \frac{(ix)^k}{k!} + \frac{1}{n!}\int_0^x (x-s)^n D^{(n+1)}(e^{is})\,ds \\ =& \sum_{k=0}^{n} \frac{(ix)^k}{k!} + \frac{i^{n+1}}{n!}\int_0^x (x-s)^n e^{is}\,ds \end{aligned} \tag{184}$$

Note that there has been an implicit use of induction in working out $D^{(n+1)}(e^{is})$. Now we have to go back and do an integration by parts as follows with $v = e^{is}$ and $dv = ie^{is}\,ds$, $du = (x-s)^n\,ds$ and $u = -\frac{(x-s)^{n+1}}{n+1}$:

$$\begin{aligned} \int_0^x (x-s)^n e^{is}\,ds =& -\frac{(x-s)^{n+1}}{n+1} e^{is}\Big|_{s=0}^{x} + \int_0^x \frac{(x-s)^{n+1}}{n+1}\, ie^{is}\,ds \\ =& \frac{x^{n+1}}{n+1} + \frac{i}{n+1}\int_0^x (x-s)^{n+1} e^{is}\,ds \end{aligned} \tag{185}$$

Now in (185) replace n with $n-1$ giving:

$$\int_0^x (x-s)^n\, e^{is}\,ds = \frac{n}{i}\Big(\int_0^x (x-s)^{n-1}\, e^{is}\,ds - \frac{x^n}{n}\Big) \tag{186}$$

Thus using (184) and (186) we have:

$$\begin{aligned} e^{ix} =& \sum_{k=0}^{n} \frac{(ix)^k}{k!} + \frac{i^{n+1}}{n!}\frac{n}{i}\Big(\int_0^x (x-s)^{n-1}\, e^{is}\,ds - \frac{x^n}{n}\Big) \\ =& \sum_{k=0}^{n} \frac{(ix)^k}{k!} + \frac{i^n}{(n-1)!}\int_0^x (x-s)^{n-1}\,(e^{is}-1)\,ds \end{aligned} \tag{187}$$

Note that with $u = x-s$ and $du = -ds$ we have that $-\frac{i^n}{(n-1)!}\int_0^x (x-s)^{n-1}\,ds = -\frac{i^n}{(n-1)!}\int_x^0 u^{n-1}\,(-du) = \frac{i^n}{(n-1)!}\int_x^0 u^{n-1}\,du = -\frac{i^n}{(n-1)!}\int_0^x u^{n-1}\,du = -\frac{i^n}{(n-1)!}\Big[\frac{u^n}{n}\Big]_0^x = -\frac{x^n}{n!}$

It follows from (187) that:

$$\Big|e^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!}\Big| \le \min\Big\{\frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!}\Big\} \tag{188}$$

To see this note that from (184) we have that, assuming $x \ge 0$ so that $u = x - s \ge 0$ for $0 \le s \le x$):

$$\begin{aligned}\Big|e^{ix}-\sum_{k=0}^{n}\frac{(ix)^k}{k!}\Big| =& \Big|\frac{i^{n+1}}{n!}\int_0^x (x-s)^n\, e^{is}\, ds\Big| \\ \leq& \frac{1}{n!}\int_0^x |x-s|^n\, ds \\ =& \frac{1}{n!}\int_x^0 u^n\,(-du) \\ & \frac{1}{n!}\int_0^x u^n\, du \\ =& \frac{x^{n+1}}{(n+1)!}\end{aligned} \tag{189}$$

If $x<0$ then we get the same result when we integrate. Hence for both cases we have that:

$$\Big|e^{ix}-\sum_{k=0}^{n}\frac{(ix)^k}{k!}\Big| \leq \frac{|x|^{n+1}}{(n+1)!} \tag{190}$$

To get the other limit we focus on (187):

$$\begin{aligned}\Big|e^{ix}-\sum_{k=0}^{n}\frac{(ix)^k}{k!}\Big| =& \Big|\frac{i^n}{(n-1)!}\int_0^x (x-s)^{n-1}\,(e^{is}-1)\, ds\Big| \\ \leq& \frac{1}{(n-1)!}\int_0^x |x-s|^{n-1}\,|e^{is}-1)\, ds \\ \leq& \frac{1}{(n-1)!}\int_0^x |x-s|^{n-1}\,(|e^{is}|+1)\, ds \\ \leq& \frac{2}{(n-1)!}\int_0^x |x-s|^{n-1}\, ds \\ =& \frac{2|x|^n}{n!}\end{aligned} \tag{191}$$

where the same logic for $x<0$ and $x\geq 0$ used above has been employed. Hence (181) is established.

Since $\phi(t)=\mathbb{E}[e^{itx}]$ and using (181) and linearity of the expectation operator we have that:

$$\Big|\phi(t)-\sum_{k=0}^{n}\frac{(it)^k}{k!}\mathbb{E}[X^k]\Big| \leq \mathbb{E}\left[\min\left\{\frac{|tX|^{n+1}}{(n+1)!},\frac{2|tX|^n}{n!}\right\}\right] \tag{192}$$

Where this gets us is that if we have any t which satisfies:

$$\lim_n \frac{|t|^n\mathbb{E}[|X|^n]}{n!}=0 \tag{193}$$

this means that:

$$\phi(t)=\sum_{k=0}^{\infty}\frac{(it)^k}{k!}\mathbb{E}[X^k] \tag{194}$$

and if $\sum_{k=0}^{\infty} \frac{|t|^k}{k!}\mathbb{E}[|X|^k] = \mathbb{E}[e^{|tX|}] < \infty$ we actually have the relationship (194) assuming X has a moment generating function over the whole line. Note that the LHS of (192) is essentially the partial sum for the infinite series and we have shown that it can be bounded by an arbitrarily small number.

If X is $N(0,1)$ then its density is $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$ and $\mathbb{E}[e^{|tX|}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i|tx|} e^{\frac{-x^2}{2}}\, dx$ but $\left| \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i|tx|} e^{\frac{-x^2}{2}}\, dx \right| \leq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |e^{i|tx|}|\, e^{\frac{-x^2}{2}}\, dx \leq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-x^2}{2}}\, dx < \infty$ so its characteristic function exists.

Billingsley [1] uses the above analysis to prove that if we have a finite k^{th} moment ie $\mathbb{E}[|X^k|] < \infty$ then we have:

$$\phi^{(k)}(0) = i^k \mathbb{E}[X^k] \tag{195}$$

where $\phi^{(k)}(t)$ is the k^{th} derivative and $\phi^{(k)}(t) = \mathbb{E}\Big[(iX)^k\, e^{itX}\Big]$ For instance, $\phi'(t) = \mathbb{E}[iX\, e^{itX}]$.

Billingsley goes on to establish the same multiplicative property that holds for moment generating functions where the random variables are independent. For instance:

$$\phi_1(t)\, \phi_2(t) = \mathbb{E}\Big[e^{it(X_1+X_2)}\Big] \tag{196}$$

I refer you to pages 346-347 and pages 388-389 of [1] for the Inversion and Uniqueness Theorem.

9 Uniform convergence proof in the context of term by term differentiation - discrete case

Recall that (37) embodies the claim that you can do term by term differentiation. The basic theorem here is that if the indexed functions are continuous and their derivatives are continuous on some interval, for example, $(-1, 1)$ in this case, then as long as the derived series is uniformly convergent you can perform the term by term differentiation.

In the probabilistic context we will obviously assume that the expectation exists ie $\exists M > 0$ such that:

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} k p_k < M \tag{197}$$

If we have functions $u_n(t)$ which are continuous and have continuous derivatives $u_n'(t)$ and $\sum_n u_n(t)$ converges to $S(t)$ while$\sum_{n=1}^{\infty} u_n'(t)$ is uniformly convergent in some interval, then in that interval $\frac{d}{dt}\left(\sum_{n=1}^{\infty} u_n(t) \right) = \sum_{n=1}^{\infty} u_n'(t)$. In this context our interval of interest is $(-1, 1)$. The first point to note is that:

$$u_k(t) = k p_k t^{k-1} \tag{198}$$

where $0 < p_k < 1$ for all k.

Because the expectation is assumed to exist we have the following estimates:

$$\begin{aligned}\sum_{k=1}^{\infty} kp_k t^{k-1} \leq & \sum_{k=1}^{\infty} \Big(\sum_{k=1}^{\infty} kp_k \Big) t^{k-1} \\ < & \sum_{k=1}^{\infty} Mt^{k-1} \\ \leq & \sum_{k=1}^{\infty} M|t|^{k-1} \\ = & \frac{M}{1-|t|}\end{aligned} \tag{199}$$

since $|t| < 1$.

Note that kp_k must be less than $\sum_{k=1}^{\infty} kp_k$.

So for any fixed $t \in (-1, 1)$, $\sum_{k=1}^{\infty} kp_k t^{k-1}$ converges.

For uniform convergence we have the following for $t \in (-1, 1)$:

$$\begin{aligned}|u_k(t)| = & kp_k |t|^{k-1} \\ \leq & kp_k \\ = & M_k\end{aligned} \tag{200}$$

But $\sum_{k=1}^{\infty} M_k = \sum_{k=1}^{\infty} kp_k < M$ which is independent of t and k. Hence uniform convergence of $\sum_{k=1}^{\infty} kp_k t^{k-1}$ is established by the Weierstrass M-test.

10 References

[1] Patrick Billingsley, Probability and Measure , Third Edition, John Wiley and Sons, 1995

[2] David Bressoud, A Radical Approach to Real Analysis , Second Edition, The Mathematical Association of America, 2007

[3] William Feller, An Introduction to Probability and Its Applications, Volume 1 , 3rd Edition, Wiley , 1968

[4] Peter Haggstrom, Basic Fourier integrals,
`https://www.gotohaggstrom.com/Basic%20Fourier%20integrals.pdf`

[5] Elias M Stein and Rami Shakarchi, Fourier Analysis: An Introduction , Princeton Lectures in Analysis 1, Princeton University Press, 2003.

[6] Peter Haggstrom, Bernoulli trials with variable probabilities - an observation by Feller , https://gotohaggstrom.com/Bernoulli%20trials%20with%20variable%20probabilities%20-%20an%20observation%20by%20Feller.pdf

11 History

Created 30 September 2023