Chapter 34

Mathematical Models Of Statistics

(34.1) Survival Analysis

Consider a population where each member will be removed from the population (failed) at some time after the 'start', for example due to failure or death. The survival time t is treat as a random variable T, which it is necessary to model.

(34.1.1) The Model

Continuous Time

For continuous time let T have probability density function $f(t)$, so that the distribution function is

$$F(t) = \int_0^\infty f(u)\,\mathrm{d}u$$

Definition 1 *The survivor function $S(t)$ for a population member is the probability that the population member has not failed at time t. Thus*

$$S(t) = \Pr\{T \geqslant t\} = 1 - F(t)$$

Definition 2 *The hazard function $h(t)$ for a population member is the probability of failure at time t, given that the population member has not already died at time t. Thus*

$$h(t) = \lim_{\delta t \to 0} \frac{\Pr\{t \leqslant T \leqslant t + \delta T \mid T \geqslant t\}}{\delta t}$$

The definition of the hazard function clearly makes sense, though its functional form is quite useless.

Lemma 3 $h(t) = \dfrac{f(t)}{S(t)}$.

Proof. Using Bayes Theorem to work with the conditional probability,

$$\begin{aligned}
\lim_{\delta t \to 0} \frac{\Pr\{t \leqslant T \leqslant t + \delta T \mid T \geqslant t\}}{\delta t} &= \lim_{\delta t \to 0} \frac{\Pr\{t \leqslant T \leqslant t + \delta t\}}{\Pr\{T \geqslant t\}\,\delta t} \\
&= \lim_{\delta t \to 0} \frac{F(t + \delta t) - F(t)}{S(t)\delta t} \\
&= \frac{1}{S(t)} \frac{\mathrm{d}F}{\mathrm{d}t} \\
&= \frac{f(t)}{S(t)}
\end{aligned}$$

□

Corollary 4 $h(t) = -\dfrac{\mathrm{d}}{\mathrm{d}t} \ln S(t)$.

Proof. Simply note that $S(t) = 1 - F(t)$ to give

$$-\frac{\mathrm{d}}{\mathrm{d}t} \ln S(t) = -\frac{\mathrm{d}}{\mathrm{d}t} \ln(1 - F(t)) = \frac{f(t)}{S(t)}$$

as required. □

Sometimes it is convenient to use the integrated hazard function, which has the obvious definition,

$$H(t) = \int_0^t h(u)\,\mathrm{d}u = -\ln S(t)$$

Discrete Time

Discrete time may occur when observations of the population are taken at certain intervals, yielding a countable set of measurement times $t_1, t_2, \ldots$. In exact analogy with continuous time,

$$\begin{aligned} f(t_i) &= \Pr\{T = t_i\} \\ S(t_i) &= \Pr\{T \geqslant t_i\} \\ h(t_i) &= \Pr\{T = t_i \mid T \geqslant t_i\} \end{aligned}$$

Clearly

$$S(t_i) = \sum_{t \geqslant t_i} f(t)$$

and using Bayes Theorem

$$h(t_i) = \frac{\Pr\{T = t_i\}}{\Pr\{T \geqslant t_i\}} = \frac{f(t_i)}{S(t_i)}$$

The following result is also available.

Lemma 5 $S(t_i) = \prod_{t<t_i} 1 - h(t)$.

Proof. First of all observe that

$$\begin{aligned} S(t_i) &= \Pr\{T \geqslant t_i\} \\ &= \Pr\{T = t_i\} + \Pr\{T = t_{i+1}\} + \ldots \\ &= f(t_i) + f(t_{i+1}) + \ldots \end{aligned}$$

Hence

$$\begin{aligned} h(t_i) &= \frac{f(t_i)}{f(t_i) + f(t_{i+1}) + \ldots} \\ 1 - h(t_i) &= \frac{f(t_{i+1}) + f(t_{i+2}) + \ldots}{f(t_i) + f(t_{i+1}) + \ldots} \\ \prod_{t<t_i} 1 - h(t) &= \left(\frac{f(t_i) + f(t_{i+1}) + \ldots}{f(t_{i-1}) + f(t_i) + \ldots}\right)\left(\frac{f(t_{i-1}) + f(t_i) + \ldots}{f(t_{i-2}) + f(t_{i-1}) + \ldots}\right)\cdots\left(\frac{f(t_2) + f(t_3) + \ldots}{f(t_1) + f(t_2) + \ldots}\right) \\ &= \frac{f(t_i) + f(t_{i+1}) + \ldots}{f(t_1) + f(t_2) + \ldots} \\ &= f(t_i) + f(t_{i+1}) + \ldots \\ &= S(t_i) \end{aligned}$$

as required. □

Corollary 6 $f(t_i) = h(t_i) \prod_{t<t_i} 1 - h(t)$.

Proof. Use the preceding Lemma with $f(t_i) = S(t_i)h(t_i)$. □

Estimation Of The Survivor Function

Trivially, the survivor function may be estimated by

$$\tilde{S}(t) = \frac{\text{Number still surviving at time } t}{\text{Size of original population}}$$

However, this assumes that the failure times of each populant has been observed. In realistic situations many of the data may censored: after some time the populant is no longer available for observation, though failure has not occurred. For example in a medical trial some subjects may cease to attend, or the trial may finish before all participants have 'failed'. This is so frequently the case that censored data cannot be ignored.

Assume that failures occur at exactly the times $t_1 < t_2 < \cdots < t_r$ but that censorings occur at some intermediate times. Let n be the population size, suppose that d_j failures occur at time t_j and let n_j be the number of remaining populants immediately before this time. Hence

$$\Pr\left\{\text{failure at time } t_j - \delta t\right\} = \frac{d_j}{n_j}$$

Hence

$$\Pr\left\{\text{survival in time interval } (t_{j-1}, t_j]\right\} = 1 - \frac{d_j}{n_j} = \frac{n_j - d_j}{n_j}$$

Hence the survivor function mat be estimated by

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{\{j|t_j<t\}} \frac{n_j-d_j}{n_j} & \text{otherwise} \end{cases}$$

This is known as the Kaplan-Meier estimator. Next its standard error is calculated.

Theorem 7 $\operatorname{sterr}\left(\hat{S}(t)\right) = \sqrt{(\hat{S}(t))^2 \sum_{\{j|t_j<t\}} \frac{d_j}{n_j - d_j}}$.

Proof. Let $\hat{p}_j = \frac{n_j-d_j}{n_j}$ then

$$\ln \hat{S}(t) = \sum_{\{j|t_j<t\}} \ln \hat{p}_j$$

$$\operatorname{var}\left(\ln \hat{S}(t)\right) = \sum_{\{j|t_j<t\}} \operatorname{var}\left(\ln \hat{p}_j\right) \tag{8}$$

Now, suppose that the number of people surviving at time t_j, given by $n_j - d_j$, is distributed as $\operatorname{Bin}(n_j, p_j)$.

Therefore

$$
\begin{aligned}
\operatorname{var}\left(n_j - d_j\right) &= n_j p_j (1 - p_j) \\
\operatorname{var} \hat{p}_j &= \frac{\operatorname{var}\left(n_j - d_j\right)}{n_j^2} \\
&= \frac{p_j(1-p_j)}{n_j} \qquad (9)
\end{aligned}
$$

Now use the Taylor series approximation

$$
\operatorname{var}(g(X)) \approx \left(\frac{\mathrm{d}g}{\mathrm{d}X}\right)^2 \operatorname{var} X
$$

So equation (9) gives

$$
\begin{aligned}
\operatorname{var} \ln \hat{p}_j &\approx \frac{1}{\hat{p}_j^2} \operatorname{var} \hat{p}_j \\
&= \frac{1}{\hat{p}_j^2} \frac{p_j(1-p_j)}{n_j} \\
&= \frac{1 - \frac{n_j - d_j}{n_j}}{n_j \frac{n_j - d_j}{n_j}} \\
&= \frac{d_j}{n_j - d_j}
\end{aligned}
$$

Using the Taylor series approximation again,

$$
\operatorname{var} \ln \hat{S}(t) \approx \frac{1}{(\hat{S}(t))^2} \operatorname{var} \hat{S}(t)
$$

Hence equation (8) gives

$$
\begin{aligned}
\frac{1}{(\hat{S}(t))^2} \operatorname{var} \hat{S}(t) &\approx \sum_{\{j|t_j<t\}} \frac{d_j}{n_j - d_j} \\
\operatorname{var} \hat{S}(t) &\approx (\hat{S}(t))^2 \sum_{\{j|t_j<t\}} \frac{d_j}{n_j - d_j} \\
\operatorname{sterr}\left(\hat{S}(t)\right) &\approx \sqrt{(\hat{S}(t))^2 \sum_{\{j|t_j<t\}} \frac{d_j}{n_j - d_j}} \qquad \square
\end{aligned}
$$

(34.1.2) Distributions & Likelihood

It is of course necessary to decide upon some distributional form for a model if it is to be used in reality. Any distribution must be defined only for $t \geqslant 0$, negative times are meaningless. One possibility is the exponential distribution which gives

$$
f(t) = \lambda e^{-\lambda t} \quad F(t) = 1 - e^{-\lambda t} \quad S(t) = E^{-\lambda t} \quad h(t) = \lambda \quad H(t) = \lambda t
$$

As the hazard function is constant, this is known as the constant hazard model. The gamma distribution, given by

$$f(t) = \frac{\beta^\alpha t^{\alpha-1} e^{-\beta t}}{\Gamma(\alpha)}$$

is a suitable distribution, though it is particularly difficult to work with. A useful alternative is the Weibull distribution, where the random variable T has $T^\gamma \sim \mathcal{E}(\lambda^\gamma)$. From defining $H(t) = (\lambda t)^\gamma$ the following can be deduced

$$h(t) = \lambda\gamma(\lambda t)^{\gamma-1} \quad S(t) = \exp\left(-(\lambda t)^\gamma\right) \quad f(t) = \lambda\gamma(\lambda t)^{\gamma-1}\exp\left(-(\lambda t)^\gamma\right)$$

Having chosen a distribution a likelihood can be calculated. Before giving a specific example, some of the calculations can be performed for a general distribution.

Let the probability density function be characterised by some parameter vector **Œ**. Populants that are censored cannot contribute a failure time to the likelihood, so instead the survivor function is used to give

$$L = \prod_{\text{uncensored}} f(t_i, \textbf{Œ}) \prod_{\text{censored}} S(t_i, \textbf{Œ})$$

where for the censored populants t_i is the first time at which censoring is noticed.

$$\begin{aligned}
l &= \sum_{\text{uncensored}} \ln f(t_i, \textbf{Œ}) + \sum_{\text{censored}} \ln S(t_i, \textbf{Œ}) \\
&= \sum_{\text{uncensored}} \ln S(t_i, \textbf{Œ}) + \sum_{\text{uncensored}} \ln h(t_i, \textbf{Œ}) + \sum_{\text{censored}} \ln S(t_i, \textbf{Œ}) \quad \text{using } h(t) = \frac{f(t)}{S(t)} \\
&= \sum_{\text{uncensored}} \ln h(t_i, \textbf{Œ}) + \sum_{\text{all}} \ln S(t_i, \textbf{Œ}) \\
&= \sum_{\text{uncensored}} \ln h(t_i, \textbf{Œ}) - \sum_{\text{all}} H(t_i, \textbf{Œ})
\end{aligned}$$

Using the constant hazard model this gives

$$l = \sum_{\text{uncensored}} \ln\lambda - \sum_{\text{all}} \lambda t_i = d\ln\lambda - \sum_{\text{all}} \lambda t_i \tag{10}$$

where d populants fail. Differentiating,

$$\hat{\lambda} = \frac{d}{\sum_{\text{all}} t_i} \qquad \frac{\partial^2 l}{\partial \lambda^2} = \frac{-d}{\lambda^2}$$

so the asymptotic variance of $\hat{\lambda}$ is $\frac{\lambda^2}{d}$. Using $\hat{\lambda}$ the survivor function and hazard function can be estimated.

Analysis Under Full Parameterisation

The probability density function, and hence S and h, may be parameterised in some general parameters **Œ** and some parameters **fi** relating to explanatory variables **z**. Hence write

$$\begin{aligned}
l(\mathbf{t}, \textbf{Œ}, \textbf{fi}) &= \sum_{\text{uncensored}} \ln f(t_i, \mathbf{z}_i, \textbf{Œ}) + \sum_{\text{censored}} \ln S(t_i, \mathbf{z}_i, \textbf{Œ}) \\
&= \sum_{\text{uncensored}} \ln h(t_i, \mathbf{z}_i, \textbf{Œ}) + \sum_{\text{all}} \ln S(t_i, \mathbf{z}_i, \textbf{Œ})
\end{aligned}$$

At this point it is necessary to specify a distribution. As an example, say f is exponentially distributed, so that the rate parameter λ is then determined by the explanatory variables **z**. Hence write $\lambda = \rho(\mathbf{z}, \textbf{fi})$ so that

$$h = \rho \qquad H = \rho t \qquad S = e^{-H}$$

and therefore

$$l(\mathbf{t}, \text{Œ}, \textbf{fi}) = \sum_{\text{uncensored}} \ln \rho(\mathbf{z}_i, \text{Œ}) - \sum_{\text{all}} \rho(\mathbf{z}_i, \text{Œ}) t_i$$

Now put $\rho(\mathbf{z}_i, \textbf{fi}) = \exp(\textbf{fi}^\top \mathbf{z}_i)$, so

$$l(\mathbf{t}, \text{Œ}, \textbf{fi}) = \sum_{\text{uncensored}} \textbf{fi}^\top \mathbf{z}_i - \sum_{\text{all}} e^{\textbf{fi}^\top \mathbf{z}_i} t_i$$

Differentiating, the expected Fisher information matrix can be obtained.

(34.1.3) Related Survival Distributions

It is often the case that two populations are under consideration, where the hazard function of one may be considered as a function of the hazard function of the other. For example in a medical trial a control group may have hazard function h_0 while a test group has a hazard function h_1. Two main relationships are considered.

Accelerated Life Model

Let population 0 have survivor function $S_0(t)$, and for population 1 put $S_1(t) = S_0(\psi t)$.

$$1 - F_1(t) = S_1(t) = S_0(\psi t) \quad \text{so} \quad f_1(t) = \psi f_0(t)$$

and similarly $h_1(t) = \psi h_0(t)$. Generally ψ will be a function of some explanatory variables $\mathbf{z}$. This model is not of much interest. The other, more interesting, relationship is discussed in the next section.

(34.1.4) Proportional Hazards Model For Related Populations

The Proportional Hazards Model

For proportional hazards, put $h_1(t) = \psi h_0(t)$, as the name suggests. Here ψ must be independent of time, but may be parameterised by explanatory variables to give $\psi = \psi(\mathbf{z})$.

For example, one may put $\psi = \exp(\textbf{fi}^\top \mathbf{z})$, which is known as the Cox model.

Generally, let λ be the hazard rate for population 0, and so let $\lambda\psi$ be the hazard rate for population 1. For constant λ *i.e.*, the constant risk model where $T \sim \mathcal{E}(\lambda)$ equation (10) gives

$$l(\psi, \lambda) = d_0 \ln \lambda - \lambda \sum_{P_0} t_i + d_0 \ln \lambda\psi - \lambda\psi \sum_{P_1} t_i$$

Differentiating gives

$$\frac{\partial l}{\partial \lambda} = \frac{d_0 + d_1}{\lambda} - \sum_{P_0} t_i - \psi \sum_{P_1} t_i \qquad \frac{\partial l}{\partial \psi} = \frac{d_1}{\psi} - \lambda \sum_{P_1} t_i$$

from which some simple algebra gives

$$\hat{\lambda} = \frac{d_0}{\sum_{P_0} t_i} \qquad \hat{\psi} = \frac{d_1 \sum_{P_0} t_i}{d_0 \sum_{P_1} t_i}$$

Finding the second partial derivatives allows the information matrix to be found, the inverse of which is the variance-covariance matrix,

$$\begin{pmatrix} \frac{-d_1}{\psi^2} & -\sum_{P_1} t_i \\ -\sum_{P_1} t_i & \frac{-(d_0+d_1)}{\lambda^2} \end{pmatrix}^{-1} = \frac{1}{(d_0+d_1)d_1 - \lambda^2\psi^2\left(\sum_{P_1} t_i\right)^2} \begin{pmatrix} (d_0+d_1)\psi^2 & -\lambda^2\psi^2\sum_{P_1} t_i \\ -\lambda^2\psi^2\sum_{P_1} t_i & d_1\lambda^2 \end{pmatrix}$$

From this the standard errors can be read off, and substituting with the estimates for λ and ψ gives

$$\begin{aligned} \text{sterr}\left(\hat{\psi}\right) &= \sqrt{\frac{(d_0+d_1)\psi^2}{(d_0+d_1)d_1 - \lambda^2\psi^2\left(\sum_{P_1} t_i\right)^2}} & \text{sterr}\left(\hat{\lambda}\right) &= \sqrt{\frac{d_1\lambda^2}{(d_0+d_1)d_1 - \lambda^2\psi^2\left(\sum_{P_1} t_i\right)^2}} \\ &= \hat{\psi}\sqrt{\frac{d_0+d_1}{d_0 d_1}} & &= \frac{\hat{\lambda}}{\sqrt{d_0}} \end{aligned}$$

Due to the skewed distributions of λ and ψ these standard errors cannot be used to find confidence intervals: the assumption of normality does not hold. However, confidence intervals for $\ln\lambda$ and $\ln\psi$ can be found in the usual way, using the Taylor series approximation to find the standard error of $\ln\lambda$ and $\ln\psi$.

General Likelihood Function

The function ψ is typically dependent on some explanatory variables $\mathbf{z}$ and there are three popular parameterisations:

- The log-linear parameterisation where $\psi(\mathbf{z}, \mathbf{fi}) = \exp \mathbf{fi}^\top \mathbf{z}$.
- The linear parameterisation where $\psi(\mathbf{z}, \mathbf{fi}) = 1 + \mathbf{fi}^\top \mathbf{z}$.
- The logistic parameterisation where $\psi(\mathbf{z}, \mathbf{fi}) = \ln\left(1 + \exp \mathbf{fi}^\top \mathbf{z}\right)$.

For whichever parameterisation is chosen it is necessary to find a likelihood function so that an estimates for $\mathbf{fi}$ and h_0 can be found.

Let the population be of size n and suppose that one populant fails at each of r failure times, $t_1 < t_2 < \cdots < t_r$. Hence there are $n - r$ censorings. Define the "risk set"

$$R(t_j) = \{\text{populants that have not failed at time } t_j\text{, or are censored just prior to } t_j.\}$$

Note that the risk set *includes* populants that fail at time t_j. For simplicity the case where there is no censoring is considered. Let $\mathbf{z}_j$ be the explanatory variables for a populant failing at time t_j then

$$\begin{aligned} &\Pr\left\{\text{populant } i \text{ fails at time } t_j \mid t_j \text{ is an observed failure time}\right\} \\ &= \Pr\left\{\text{populant with explanatory variables } \mathbf{z}_j \text{ failes at time } t_j \mid \text{one failure occurs at each } t_j\right\} \\ &= \frac{\Pr\left\{\text{populant with explanatory variables } \mathbf{z}_j \text{ failes at time } t_j\right\}}{\Pr\left\{\text{one failure occurs at each } t_j\right\}} \\ &= \frac{h_i(t_j)}{\sum_{k\in R(t_j)} h_k(t_j)} \\ &= \frac{\psi(i)}{\sum_{k\in R(t_j)} \psi(k)} \end{aligned}$$

and hence the likelihood function is

$$L = \prod_{j=1}^{r} \frac{\psi(i)}{\sum_{k\in R(t_j)} \psi(k)}$$

For censored data it is assumed that censorings occur immediately after failure times. Censorings in the time interval (t_{j-1}, t_j) effect the risk set $R(t_j)$. The same likelihood function is used, though it becomes only a partial likelihood function.

Derivatives Of The General Likelihood Function

Write $\psi(i)$ for $\psi(\mathbf{fi}, \mathbf{z}_i)$ and $\psi_u(i)$ for the derivative of $\psi(\mathbf{fi}, \mathbf{z}_i)$ with respect to the uth element of $\mathbf{fi}$. Let D be the set of populants that fail and are not censored. Hence

$$l = \sum_{i \in D} \left(\ln \psi(i) - \ln \sum_{k \in R(t_j)} \psi(k) \right)$$

Let l_i be the ith term of this sum then

$$\begin{aligned}
\frac{\partial l_i}{\partial \beta_u} &= \frac{\psi_u(i)}{\psi(i)} - \frac{\sum_{k \in R(t_j)} \psi_u(i)}{\sum_{k \in R(t_j)} \psi(k)} \\
\frac{\partial^2 l_i}{\partial \beta_u \partial \beta_v} &= \frac{\psi_{uv}(i)}{\psi(i)} - \frac{\psi_u(i)\psi_v(i)}{(\psi(i))^2} - \frac{\sum_{k \in R(t_j)} \psi_{uv}(i)}{\sum_{k \in R(t_j)} \psi(k)} + \frac{\sum_{k \in R(t_j)} \psi_u(i)\psi_v(i)}{\left(\sum_{k \in R(t_j)} \psi(k)\right)^2}
\end{aligned}$$

A covariance matrix, and note putting $\psi(\mathbf{z}, \mathbf{fi}) = \exp \mathbf{fi}^\top \mathbf{z}$ gives a bit of a simplification.

Multiple failures at a single time

(34.2) Mathematical Foundation Of Statistics

(34.2.1) Probability Space

As with many branches of mathematics, sets play a fundamental rôle in the underlying workings of statistics. In statistics it is usual to consider some set Ω of 'outcomes', and take a family of subsets $\mathcal{A} \subseteq 2^\Omega$ as 'events'.

Sets

A sequence of sets $\{A_n\}_{n=1}^\infty$ is monotone increasing if $A_n \subseteq A_{n+1}$ for all n. In this case the definition

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

makes sense. Similarly, for a monotone decreasing sequence where $A_n \supseteq A_{n+1}$ define

$$\lim_{n \to \infty} A_n = \bigcap_{n=1}^\infty A_n$$

For an arbitrary sequence define

$$\varliminf_{n \to \infty} A_n = \lim_{n \to \infty} \inf_{k \geqslant n} A_k = \lim_{n \to \infty} \bigcap_{k \geqslant n} A_k \quad \text{and} \quad \varlimsup_{n \to \infty} A_n = \lim_{n \to \infty} \sup_{k \geqslant n} A_k = \lim_{n \to \infty} \bigcup_{k \geqslant n} A_k$$

If both of these limits exist and are the same, A say, then write $\lim A_n = A$. It can be shown that in general $\varliminf A_n \subseteq \varlimsup A_n$, and equality need not hold.

Fields

Definition 11 *A non-empty family of sets $\mathcal{A}$ that is closed under finite unions and under complementation is called a field.*

Equivalently, the requirement of closure under finite unions may be replaced with closure under finite intersections, as the following lemma shows.

Lemma 12 *A non-empty family of sets $\mathcal{A}$ that is closed under complementation is closed under finite unions if and only if it is closed under finite intersections.*

Proof. Let $\mathcal{A}$ be a field. Let $A_1, A_2, \ldots, A_n \in \mathcal{A}$ then $A_i^{\complement} \in \mathcal{A}$ for all i.

- ($\Rightarrow$) Since $\mathcal{A}$ is closed under finite unions, $\bigcup_{i=1}^{n} A^{\complement} \in \mathcal{A}$ and therefore $\left(\bigcup_{i=1}^{n} A^{\complement}\right)^{\complement} \in \mathcal{A}$. But then by De-Morgan's laws, $\bigcap_{i=1}^{n} a_i \in \mathcal{A}$.
- ($\Leftarrow$) Since $\mathcal{A}$ is closed under finite intersections, $\bigcap_{i=1}^{n} A^{\complement} \in \mathcal{A}$ and therefore $\left(\bigcap_{i=1}^{n} A^{\complement}\right)^{\complement} \in \mathcal{A}$. But then by De-Morgan's laws, $\bigcup_{i=1}^{n} a_i \in \mathcal{A}$. □

The family $\mathcal{A}$ must have some underlying set of which its elements are subsets, Ω say. Therefore $\mathcal{A} \subseteq 2^{\Omega}$. Every field contains both $\emptyset$ and Ω, indeed $\{\emptyset, \Omega\}$ is a field.

Fields are not sufficiently 'well behaved', and so a weakening of the definition is permitted.

Definition 13 *A non-empty family of sets $\mathcal{A}$ that is closed under complementation and countable unions is called a σ-field.*

Once again, replacing "union" with "intersection" gives a completely equivalent definition. Clearly a σ-field is a field.

A particularly useful σ-field is the Borel field, $\mathcal{B}$, which is the field generated from

$$\mathcal{C} = \{(x, \infty) \mid x \in \mathbb{R}\}$$

Probability Space

Definition 14 *Probability space is an ordered triplet $(\Omega, \mathcal{F}, \Pr)$ where Ω is a non-empty set, $\mathcal{F}$ is a σ-field of subsets of Ω, and $\Pr\colon \mathcal{F} \to [0,1]$ is the probability measure.*

A probability space has the following important properties. They are not difficult to prove, but the proof is omitted here. Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space and let $A, B, A_1, A_2, \ldots, A_n, \cdots \in \mathcal{F}$.

1. $\Pr\{A\} \geqslant 0$
2. $\Pr\{\Omega\} = 1$
3. If $A = \bigcup_{i=1}^{n} A_i$ is a disjoint union then $\Pr\{A\} = \sum_{i=1}^{n} \Pr\{A_i\}$.
4. If $A = \bigcup_{i=1}^{\infty} A_i$ is a disjoint union then $\Pr\{A\} = \sum_{i=1}^{\infty} \Pr\{A_i\}$.
5. $\Pr\{\emptyset\} = 0$.
6. If $A \subseteq B$ then $\Pr\{A\} \leqslant \Pr\{B\}$.
7. If $A = \bigcup_{i=1}^{n} A_i$ then $\Pr\{A\} \leqslant \sum_{i=1}^{n} \Pr\{A_i\}$.

The probability measure is a continuous function in the sense that if $A_n \to A$ as $n \to \infty$ then $\Pr\{A_n\} \to \Pr\{A\}$.

Theorem 15 *Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space and let $\{A_i\}_{i=1}^{\infty}$ be a sequence in $\mathcal{F}$ with limit A. If $A_n \to A$ as $n \to \infty$ then $\Pr\{A_n\} \to \Pr\{A\}$.*

Proof. First of all suppose that $\{A_n\}$ is increasing, and write $B_n = A_{n+1} \setminus A_n$ so that $\{B_n\}$ is a collection of disjoint elements of $\mathcal{F}$. Now, $A_n = \bigcup_{i=1}^{n} B_i$ and furthermore $\bigcup_{i=1}^{\infty} B_i = A$. Hence $\Pr\{A_n\} = \sum_{i=1}^{\infty} \Pr\{B_i\}$ and

$$\lim_{n\to\infty} \Pr\{A_n\} = \lim_{n\to\infty} \sum_{i=1}^{n} \Pr\{B_i\} = \sum_{i=1}^{\infty} \Pr\{B_i\} = \Pr\{A\} \tag{16}$$

Similarly if $\{A_n\}$ is decreasing.

Consider now some arbitrary sequence $\{A_n\}$ then

$$\begin{aligned} \bigcap_{k\geqslant n} A_k \subseteq \qquad & A_n \subseteq \bigcup_{k\geqslant n} A_k \\ \Rightarrow \quad \Pr\left\{\bigcap_{k\geqslant n} A_k\right\} \leqslant & \Pr\{A_n\} \leqslant \Pr\left\{\bigcup_{k\geqslant n} A_k\right\} \end{aligned} \tag{17}$$

But since $A_n \to A$ as $n \to \infty$,

$$\lim_{n\to\infty} \bigcap_{k\geqslant n} A_k = \varliminf_{n\to\infty} A_n = A \quad \text{and} \quad \lim_{n\to\infty} \bigcup_{k\geqslant n} A_k = \varlimsup_{n\to\infty} A_k = A$$

Using equation (16) on the right hand inequality in equation (17) and the corresponding equation for decreasing sequences on the left hand inequality, the squeeze rule shows that $\Pr\{A_n\} \to \Pr\{A\}$ as $n \to \infty$. □

(34.2.2) Random Variables

Random Variables

Definition 18 *Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space. A function $X\colon \Omega \to \mathbb{R}$ is a random variable if for any $E \in \mathcal{B}$, $X^{-1}(E) \in \mathcal{F}$.*

This apparently peculiar property allows X to induce a probability measure P_X on $\mathcal{B}$ by $P_X(E) = \Pr\{X^{-1}(E)\}$. This gives a new probability space $(\mathbb{R}, \mathcal{B}, P_X)$.

Definition 19 *Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space, $X\colon \Omega \to \mathbb{R}$ be a random variable, and P_X be the measure induced by X. The function*

$$F_X\colon \mathbb{R} \to [0,1] \quad \textit{defined by} \quad F_X\colon x \mapsto P_X\{(-\infty, x)\} = \Pr\{\omega \in \Omega \mid X(\omega) \leqslant x\}$$

is called the distribution of X.

The distribution function is a right-continuous, non-decreasing function. It has limit 0 at $-\infty$ and limit 1 at ∞. Moreover, any function with these four properties is a distribution function of some random variable.

The derivative (Ralon-Nikodym derivative) of P_X is called the probability density function.

Note that if $g\colon \mathbb{R} \to \mathbb{R}$ and X is a random variable then $g(X)$ is also a random variable.

Definition 20 *Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space, $X\colon \Omega \to \mathbb{R}$ be a random variable, and $g\colon \mathbb{R} \to \mathbb{R}$. If $g(X)$ is integrable over Ω then the expectation of $g(X)$ is*

$$\begin{aligned}\mathbb{E}\, g(X) &= \int_\Omega g(X(\omega))\, \mathrm{d}\Pr\{\omega\} \\ &= \int_{\mathbb{R}} g(x)\, \mathrm{d}P_X \\ &= \int_{\mathbb{R}} g(x)\, \mathrm{d}F(x)\end{aligned}$$

Theorem 21 (Hölder Inequality) *Let $p, q \in \mathbb{R}^+$ with $1 < p < q$ and $\frac{1}{p} + \frac{1}{q} = 1$. If X and Y are random variables such that $\mathbb{E}\,|X|^p$ and $\mathbb{E}\,|Y|^q$ exist then*

$$\mathbb{E}\,|XY| \leqslant (\mathbb{E}\,|X|^p)\,\frac{1}{p}\,(\mathbb{E}\,|Y|^q)\,\frac{1}{q}$$

Theorem 22 (Minkowski Inequality) *Let $p \in \mathbb{R}$ with $p \geqslant 1$ and let X and Y be random variables such that $\mathbb{E}\,X^p$ and $\mathbb{E}\,Y^p$ exist. Then*

$$(\mathbb{E}\,(X+Y)^p)^{\frac{1}{p}} \leqslant (\mathbb{E}\,X^p)^{\frac{1}{p}} + (\mathbb{E}\,Y^p)^{\frac{1}{p}}$$

Convergence

Definition 23 (Almost Sure Convergence) *Let $\{X_n\}_{n=1}^\infty$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, \Pr)$. $\{X_n\}$ converges to the random variable X almost surely if and only if*

$$\forall A \in \mathcal{F} \text{ with } \Pr\{A\} = 0 \quad \forall \omega \in A^{\complement} \quad |X_n(\omega) - X(\omega)| \to 0 \text{ as } n \to \infty$$

In this case write $X_n \xrightarrow{as} X$ as $n \to \infty$.

Roughly speaking, this means that the convergence need only hold for those $\omega in \Omega$ that can occur: If $\omega \in \Omega$ belongs to no element $A \in \mathcal{F}$ with $\Pr\{A\} > 0$ then the convergence need not hold for this ω since it almost surely will never occur.

Almost sure convergence is not easy to check, indeed it is a very strong form of convergence as is demonstrated later in Theorem 27. Fortunately there is a criterion for almost sure convergence.

Theorem 24 *Let $\{X_n\}_{n=1}^\infty$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, \Pr)$. If*

$$\sum_{n=1}^\infty \Pr\{|X_n - X| \geqslant \varepsilon\}$$

exists (i.e., *is finite) for all $\varepsilon > 0$ then $X_n \xrightarrow{as} X$ as $n \to \infty$.*

Proof. Now,

$$\Pr\left\{\bigcup_{k=n}^\infty \{\omega \in \Omega \mid |X_k(\omega) - X(\omega)| \geqslant \varepsilon\}\right\} \leqslant \sum_{k=n}^\infty \Pr\{|X_k - X| \geqslant \varepsilon\}$$

The term on the right here is the 'tail end' of a sum which, by hypothesis, converges. Hence

$$\lim_{n\to\infty} \sum_{k=n}^\infty \Pr\{|X_k - X| \geqslant \varepsilon\} = 0$$

and so by the squeeze rule the result is shown. □

Definition 25 (Convergence In Probability) *Let $\{X_n\}_{n=1}^\infty$ be a sequence of random variables on a probability space*

$(\Omega, \mathcal{F}, \Pr)$. *$\{X_n\}$ converges to the random variable X in probability if*

$$\forall \varepsilon > 0 \quad \Pr\{\omega \mid |X_n(\omega) - X(\omega)| \geqslant \varepsilon\} \to 0 \text{ as } n \to \infty$$

In this case write $X_n \xrightarrow{p} X$ as $n \to \infty$.

Theorem 26 *If $X_n \xrightarrow{p} X$ then $|X_n - X_m| \xrightarrow{p} 0$ as $n, m \to \infty$.*

Proof. Define the sets

$$A = \{\omega \mid |X_n(\omega) - X_m(\omega)| \geqslant \varepsilon\}$$
$$B = \{\omega \mid |X_n(\omega) - X(\omega)| \geqslant \frac{\varepsilon}{2}\} \cup \{\omega \mid |X_m(\omega) - X(\omega)| \geqslant \frac{\varepsilon}{2}\}$$

By the triangle inequality

$$|X_n(\omega) - X(\omega)| + |X_m(\omega) - X(\omega)| \leqslant |X_n(\omega) - X_m(\omega)| \leqslant \varepsilon$$

and so one of $|X_n(\omega) - X(\omega)|$ and $|X_m(\omega) - X(\omega)|$ must be at least $\frac{\varepsilon}{2}$. Hence $A \subseteq B$. Hence

$$\begin{aligned} \Pr\{|X_n - X_m \geqslant \varepsilon\} &\leqslant \Pr\left\{\{|X_n - X| \geqslant \frac{\varepsilon}{2}\} \cup \{|X_m - X| \geqslant \frac{\varepsilon}{2}\}\right\} \\ &\leqslant \Pr\left\{|X_n - X| \geqslant \frac{\varepsilon}{2}\right\} + \Pr\left\{|X_m - X| \geqslant \frac{\varepsilon}{2}\right\} \qquad \to 0 \text{ as } n, m \to \infty \end{aligned}$$

Hence by the squeeze rule the result is shown. □

Theorem 27 *If $X_n \xrightarrow{as} X$ as $n \to \infty$ then $X_n \xrightarrow{p} X$ as $n \to \infty$.*

Proof. Well, $X_n \xrightarrow{\text{as}} X$ if and only if

$$\forall \varepsilon > 0 \lim_{n \to \infty} \Pr\left\{\bigcup_{k=n}^{\infty} \{|X_k - X| > \varepsilon\}\right\} = 0$$

Now,

$$\{|X_n - X > \varepsilon\} \subseteq \bigcup_{k=n}^{\infty} \{|X_k - X| > \varepsilon\}$$
$$\text{so } \Pr\{|X_n - X > \varepsilon\} \leqslant \Pr\left\{\bigcup_{k=n}^{\infty} \{|X_k - X| > \varepsilon\}\right\}$$

The right hand side of this tends to 0 as $n \to \infty$ and hence by the squeeze rule the result is shown. □

Definition 28 *Let X be a random variable. If $\mathbb{E}|X|^p$ exists* (i.e., *is finite) then write $X \in L_p$.*

Definition 29 (Convergence In p-th Order Mean) *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, \Pr)$ such that $X_n \in L_p$. $\{X_n\}$ converges to the random variable X in p-th order mean if*

$$\mathbb{E}|X_n - X|^p \to 0 \text{ as } n \to \infty$$

In this case write $X_n \xrightarrow{L_p} X$ as $n \to \infty$.

Theorem 30 *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables with $X_n \in L_p$. If $X_n \xrightarrow{L_p} X$ as $n \to \infty$ then $\mathbb{E}|X|^p \to \mathbb{E}|X|^p$ as $n \to \infty$ for any $p > 0$.*

Proof. If $p \geqslant 1$ then Theorem 22 applies, so that for random variables $U, V \in L_p$

$$(\mathbb{E}|U + V|^p)^{\frac{1}{p}} \leqslant (\mathbb{E}|U|^p)^{\frac{1}{p}} + (\mathbb{E}|V|^p)^{\frac{1}{p}}$$

Put $U = X_n - X$ and $V = X$ to give

$$(\mathbb{E}\,|X_n|^p)^{\frac{1}{p}} \leqslant (\mathbb{E}\,|X_n - X|^p)^{\frac{1}{p}} + (\mathbb{E}\,|X|^p)^{\frac{1}{p}}$$

Now putting $U = X_n - X$ and $V = X_n$ gives

$$(\mathbb{E}\,|X|^p)^{\frac{1}{p}} \leqslant (\mathbb{E}\,|X_n - X|^p)^{\frac{1}{p}} + (\mathbb{E}\,|X_n|^p)^{\frac{1}{p}}$$

and hence

$$(\mathbb{E}\,|X|^p)^{\frac{1}{p}} - (\mathbb{E}\,|X_n - X|^p)^{\frac{1}{p}} \leqslant (\mathbb{E}\,|X_n|^p)^{\frac{1}{p}} \leqslant (\mathbb{E}\,|X_n - X|^p)^{\frac{1}{p}} + (\mathbb{E}\,|X|^p)^{\frac{1}{p}} \qquad \square$$

which gives the required result.

If $0 < p < 1$, please complete this proof

Theorem 31 *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables with $X_n \in L_p$. If $X_n \xrightarrow{L_p} X$ as $n \to \infty$ then $X_n \xrightarrow{p} X$ as $n \to \infty$.*

Proof. By problem sheet 1, question 5…

For any $\varepsilon > 0$

$$\Pr\{|X_n - X| \geqslant \varepsilon\} \leqslant \frac{\mathbb{E}\,|X_n - X|^p}{\varepsilon^p}$$

But since $X_n \xrightarrow{L_p} X$ as $n \to \infty$ the squeeze rule shows that $X_n \xrightarrow{p} X$ as $n \to \infty$. $\square$

If also the random variables X_n are bounded almost surely, then the converse of this theorem holds.

Definition 32 (Convergence In Distribution (Law)) *Let $\{F_n(x)\}_{n=1}^{\infty}$ be a sequence of distribution functions. If $F_n(x) \to F(x)$ as $n \to \infty$ for all continuity points x of F then the sequence $\{F_n\}$ converges weakly (or in law) to F.*

If X_n has distribution function F_n that converges in law to F then write $X_n \xrightarrow{\mathcal{L}} X$. In fact F need not be a distribution function, so this only makes sense when X has distribution function F.

A subtle difference between convergence in law and the other varieties of convergence is that the X_n need not be defined on the same probability space: the conditions for convergence are couched entirely in terms of the distribution function, which 'hides' this information.

Theorem 33 *If $X_n \xrightarrow{p} X$ as $n \to \infty$ then $X_n \xrightarrow{\mathcal{L}} X$ as $n \to \infty$.*

Proof. Let $x, x' \in \mathbb{R}$ then

$$\begin{aligned} F(x') &= \Pr\{X \leqslant x'\} \\ &= \Pr\{X_n \leqslant x \text{ and } X \leqslant x'\} + \Pr\{X_n > x \text{ and } X \leqslant x'\} \\ &\leqslant \Pr\{X_n \leqslant x\} + \Pr\{X_n > x \text{ and } X \leqslant x'\} \\ &= F_n(x) + \Pr\{X_n > x \text{ and } X \leqslant x'\} \\ &\leqslant F_n(x) + \Pr\{|X_n - X| \geqslant x' - x\} \ \text{ when } x' < x \end{aligned} \tag{34}$$

Considering now some $x'' > x$, by symmetry with the above case

$$F_n(x) \leqslant F(x'') + \Pr\{|X_n - X| \geqslant x'' - x\} \tag{35}$$

Using equation (34) and equation (35) gives

$$F(x') - \Pr\left\{|X_n - X| \geqslant x - x'\right\} \leqslant F_n(x) \leqslant F(x'') - \Pr\left\{|X_n - X| \geqslant x'' - x\right\}$$

Now let $x' \to x^-$ and $x'' \to x^+$ to give

$$\lim_{n\to\infty} F_n(x) = F(x)$$

i.e., $X_n \xrightarrow{\mathcal{L}} X$ as $n \to \infty$. □

Theorem 36 *Let c be a constant. Then $X_n \xrightarrow{p} c$ as $n \to \infty$ if and only if $X_n \xrightarrow{\mathcal{L}} c$ as $n \to \infty$.*

Proof. First of all it is necessary to find the distribution function of the random variable X that has $X(\omega) = c$ for all $\omega \in \Omega$. Recall for a probability space $(\Omega, \mathcal{F}, \Pr)$ that X is a function, in this case

$$X\colon \Omega \to \mathbb{R} \quad \text{defined by} \quad X\colon \omega \mapsto c$$

Now let $E \subseteq \mathbb{R}$ so that $E \in \mathcal{C}$. Then

$$X^{-1}(E) = \begin{cases} \varnothing & \text{if } c \notin E \\ \Omega & \text{if } c \in E \end{cases} \quad \text{so} \quad F(x) = P_X((-\infty, x]) = \begin{cases} 0 & \text{if } x < c \\ 1 & \text{if } x \geqslant c \end{cases}$$

($\Rightarrow$) Apply Theorem 33 with $F(x)$ as calculated above.

($\Leftarrow$) Suppose that F_n is convergent to F (as calculated above) in law, then

$$\begin{aligned} \Pr\{|X_n - c| \geqslant \varepsilon\} &= \Pr\{X_n \geqslant c + \varepsilon\} + \Pr\{X_n \leqslant c - \varepsilon\} \\ &= 1 - \Pr\{X_n < c + \varepsilon\} + \Pr\{X_n \leqslant c - \varepsilon\} \\ &= 1 - F_n(c + \varepsilon) + F_n(c - \varepsilon) \end{aligned} \tag{37}$$

Now, as $n \to \infty$, $F_n \to F$. Furthermore,

$$\lim_{\varepsilon\to 0} F(c + \varepsilon) = 1 \quad \text{and} \quad \lim_{\varepsilon\to 0} F(c - \varepsilon) = 0$$

so that equation (37) gives the required result. □

(34.2.3) Laws Of Large Numbers

Consider a sequence of random variables $\{X_i\}_{i=1}^n$ that are independently and identically distributed. A "law of large numbers" is a condition under which the average $\frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^n X_i$ is convergent to a constant.

Theorem 38 (Chebyshev's Weak Law Of Large Numbers) *Let $\{X_i\}_{i=1}^n$ be a sequence of independently and identically distributed random variables such that $\mathbb{E}\, X_n^2$ exists. If $\exists \gamma > 0$ such that $\operatorname{var} X_i \leqslant \gamma$ for all i then*

$$\frac{S_n - \mathbb{E}\, S_n}{n} \xrightarrow{p} 0 \text{ as } n \to \infty$$

Theorem 39 (Khintchine's Weak Law Of Large Numbers) *Let $\{X_i\}_{i=1}^n$ be a sequence of independently and identically distributed random variables each with $\mathbb{E}\, X_i = \mu$ for some constant μ. Then*

$$\frac{S_n}{n} \xrightarrow{p} \mu \text{ as } n \to \infty$$

Theorem 40 (Strong Law Of Large Numbers) *Let $\{X_i\}_{i=1}^n$ be a sequence of independently and identically distributed random variables each with $\mathbb{E}\, X_i = 0$ and $\operatorname{var} X_i = \sigma_i^2$ (finite). If*

$$\sum_{i=1}^{\infty} \frac{\sigma_i^2}{n}$$

exists (i.e., *is finite) then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{S_n}{n} \xrightarrow{as} 0 \text{ as } n \to \infty$$

Note that equivalently, if $\mathbb{E}\, X_i = \mu_i$ then with the same condition on the variances this result becomes

$$\frac{1}{n}\sum_{i=1}^{n} (X_i - \mu_i) \xrightarrow{\text{as}} 0 \text{ as } n \to \infty$$

(34.3) Mathematics Of Discrete Probability Distributions

(34.3.1) Discrete Probability Distributions

A discrete probability distribution refers to a random variable X taking values on a set $\{0, 1, \dots\}$ with probabilities $\Pr\{X = x\} = p_x$. It is required that $p_x \geqslant 0$ and that

$$\sum_{x=0}^{\infty} p_x = 1$$

Hence discrete probability distributions are sequences of real numbers which sum to 1. Throughout this section, three important examples will be considered: the Poisson distribution, the binomial distribution, and the negative binomial distribution.

The Poisson distribution is constructed from the exponential series

$$\begin{aligned} e^{\lambda} &= 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots \\ 1 &= e^{-\lambda} + e^{-\lambda}\lambda + \frac{e^{-\lambda}\lambda^2}{2!} + \frac{e^{-\lambda}\lambda^3}{3!} + \dots \end{aligned}$$

so that $p_x = \frac{e^{-\lambda}\lambda^x}{x!}$ for $x \in \mathbb{N} \cup \{0\}$ and $\lambda \in \mathbb{R}^+$.

Similarly, the binomial distribution arrises from the binomial series, where for $0 < p < 1$

$$\left(1 + \frac{p}{1-p}\right)^n = 1 + n\left(\frac{p}{1-p}\right) + \binom{n}{2}\left(\frac{p}{1-p}\right)^2 + \dots + \binom{n}{x}\left(\frac{p}{1-p}\right)^x + \dots + \left(\frac{p}{1-p}\right)^n$$

so that for $x \in \{0, 1, \dots, n\}$

$$\begin{aligned} p_x &= \binom{n}{x}\left(\frac{p}{1-p}\right)^x \left(1 + \frac{p}{1-p}\right)^{-n} \\ &= \binom{n}{x}\left(\frac{p}{1-p}\right)^x (1-p)^n \\ &= \binom{n}{x} p^x (1-p)^{n-x} \end{aligned}$$

Of course the negative binomial distribution arrises from the negative binomial series, where for $0 < p < 1$

$$(1-(1-p))^{-n} = 1 + n(1-p) + \frac{n(n-1)}{2!}(1-p)^2 + \cdots + \binom{n+x-1}{x}(1-p)^x + \ldots$$

so that for $x \in \mathbb{N} \cup \{0\}$

$$p_x = \binom{n+x-1}{x}(1-p)^x p^n$$

(34.3.2) Probability Generating Functions

A probability generating function Π for a discrete probability distribution p_x with $x \in \mathbb{N} \cup \{0\}$ is the power series

$$\Pi(z) = \sum_{x=0}^{\infty} p_x z^x$$

Notice that for $|z| < 1$ the series is always convergent, though is particular cases it may converge for a larger range of z. Also, $\Pi(1) = 1$.

In fact, given any function $\Pi(z)$ such that $\Pi(1) = 1$ and all derivatives evaluated at 0 (exist and) are positive, and a value for $\Pi(0)$, a probability generating function can be constructed as

$$\Pi(z) = \Pi(0) + \left.\frac{\mathrm{d}\Pi}{\mathrm{d}z}\right|_{z=0} z + \left.\frac{\mathrm{d}^2\Pi}{\mathrm{d}z^2}\right|_{z=0} \frac{z^2}{2!} + \cdots + \left.\frac{\mathrm{d}^x\Pi}{\mathrm{d}z^x}\right|_{z=0} \frac{z^x}{x!} + \ldots$$

Hence there is a 1-to-1 correspondence between probability generating functions and discrete probability distributions.

For the Poisson distribution

$$\begin{aligned}
\Pi(z) &= \sum_{x=0}^{\infty} z^x p_x \\
&= \sum_{x=0}^{\infty} z^x \frac{e^{-\lambda}\lambda^x}{x!} \\
&= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda z)^x}{x!} \\
&= e^{-\lambda} e^{\lambda z} = e^{-\lambda(1-z)}
\end{aligned}$$

Similarly, for the binomial and negative binomial distributions,

$$\begin{aligned}
\Pi(z) &= \sum_{x=0}^{n} z^x \binom{n}{x} p^x (1-p)^{n-x} \\
&= (1-p)^n \sum_{x=0}^{n} \binom{n}{x} \left(\frac{pz}{1-p}\right)^n \\
&= \frac{1}{(1-p)^n} \left(1 + \frac{pz}{1-p}\right)^n \\
&= (1-p+zp)^n
\end{aligned}$$

$$\begin{aligned}
\Pi(z) &= \sum_{x=0}^{\infty} z^n \binom{n+x-1}{x} (1-p)^x p^n \\
&= p^n \sum_{x=0}^{\infty} \binom{n+x-1}{x} (z(1-p))^x \\
&= \left(\frac{p}{1-z(1-p)}\right)^n
\end{aligned}$$

(34.3.3) Obtaining Moments From Probability Generating Functions

The Mean

The mean of a probability distribution is readily obtained from the probability generating function as follows.

$$
\begin{aligned}
\mu &= \sum_{x=0}^{\infty} x p_x \\
&= \sum_{x=0}^{\infty} \left.\frac{\mathrm{d}}{\mathrm{d}z} z^x \right|_{z=1} p_x \\
&= \left.\frac{\mathrm{d}}{\mathrm{d}z} \left(\sum_{x=0}^{\infty} p_x z^x \right) \right|_{z=1} \\
&= \left.\frac{\mathrm{d}\Pi}{\mathrm{d}z}\right|_{z=1}
\end{aligned}
$$

The Variance

The variance is obtained from the probability generating function through a rather more convoluted process. First of all, observe the following

$$
\begin{aligned}
\sigma^2 &= \mathbb{E}\,(X-\mu)^2 \\
&= \mathbb{E}\,X^2 - 2\mu\,\mathbb{E}\,X + \mu^2 \\
&= \mathbb{E}\,X^2 - \mu^2 \\
&= \mathbb{E}\,X(X-1) + \mathbb{E}\,X - \mu^2 \\
&= \mathbb{E}\,X(X-0) + \mu - \mu^2
\end{aligned}
$$

Now a similar trick as with the mean can be used

$$
\begin{aligned}
\sigma^2 &= \sum_{x=0}^{\infty} x(x-1)p_x + \mu - \mu^2 \\
&= \sum_{x=0}^{\infty} \left.\frac{\mathrm{d}^2}{\mathrm{d}z^2} z^x \right|_{z=1} + \mu - \mu^2 \\
&= \left.\frac{\mathrm{d}^2}{\mathrm{d}z^2} \left(\sum_{x=0}^{\infty} z^x p_x \right) \right|_{z=1} + \mu - \mu^2 \\
&= \left.\frac{\mathrm{d}^2\Pi}{\mathrm{d}z^2}\right|_{z=1} + \mu - \mu^2
\end{aligned}
$$

Of course this process may be extended to give

$$
\mathbb{E}\left(X(X-1)(X-2)\ldots(X-n)\right) = \left.\frac{\mathrm{d}^{n+1}\Pi}{\mathrm{d}z^{n+1}}\right|_{z=1}
$$

Examples

Of course, this is readily applicable to the Poisson distribution, binomial distribution, and negative binomial distribution. For the Poisson distribution $\Pi(z) = \exp\left(-\lambda(1-z)\right)$ and so

$$\begin{aligned} \mu &= \left.\frac{d\Pi}{dz}\right|_{z=1} \\ &= \left.\frac{d}{dz}e^{-\lambda(1-z)}\right|_{z=1} \\ &= \lambda e^{-\lambda(1-1)} \\ &= \lambda \end{aligned} \qquad \begin{aligned} \sigma^2 &= \left.\frac{d^2\Pi}{dz^2}\right|_{z=1} + \mu - \mu^2 \\ &= \left.\frac{d^2}{dz^2}e^{-\lambda(1-z)}\right|_{z=1} + \mu - \mu^2 \\ &= \lambda^2 e^{-\lambda(1-1)} + \mu - \mu^2 \\ &= \lambda \end{aligned}$$

Variance-Mean Relationships

Frequently there is a relationship between the variance and the mean of a probability distribution, and it can be of interest to note this.

- For the Poisson distribution $\sigma^2 = \mu$.
- For the binomial distribution $\sigma^2 = \mu\left(1 - \frac{\mu}{n}\right)$.
- For the negative binomial distribution $\sigma^2 = \mu\left(\frac{\mu}{n} + 1\right)$.

Notice that the binomial distribution has its variance less than its mean: it is under-dispersed. Similarly, the negative binomial distribution is over dispersed. Notice also that as $n \to \infty$ both of these approach the Poisson distribution.

(34.3.4) Mixed Distributions

The three distributions covered in detail so far doe not offer much flexibility. A general way to construct over dispersed or under dispersed distributions is needed.

To construct an over dispersed distribution, take one of the three well known distributions and let the parameter of the distribution vary between observations. Trivially this leads to over dispersion.

Mixed Poisson Distribution

Consider the Poisson distribution with parameter λ which itself has a Poisson distribution. Let λ be an observation of the Poisson random variable Λ, then

$$\Pr\{X = x \mid \Lambda = \lambda\} = \frac{e^{-\lambda}\lambda^x}{x!}$$

The conditioning is removed in the usual way

$$\begin{aligned} (\lambda\text{discrete}) \quad \Pr\{X = x\} &= \sum_{\lambda=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!}\Pr\{\Lambda = \lambda\} \\ (\lambda\text{continuous}) \quad \Pr\{X = x\} &= \int_0^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} f(\lambda)\, d\lambda \end{aligned}$$

The discrete case usually has to be done numerically. However, the continuous case can be pursued by considering the probability generating function.

$$
\begin{aligned}
\Pi(z) &= \sum_{x=0}^{\infty} z^x \Pr\{X = x\} \\
&= \sum_{x=0}^{\infty} z^x \int_0^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} f(\lambda)\, \mathrm{d}\lambda \\
&= \int_0^{\infty} e^{-\lambda} f(\lambda) \sum_{x=0}^{\infty} \frac{(\lambda z)^x}{x!}\, \mathrm{d}\lambda \\
&= \int_0^{\infty} e^{-\lambda} f(\lambda) e^{-\lambda z} \\
&= \int_0^{\infty} e^{-\lambda(1-z)} f(\lambda)\, \mathrm{d}\lambda \\
&= f^*(1-z)
\end{aligned}
$$

where f^* denoted the Laplace transform of f.

Definition 41 *The Laplace transform of a function f is $\mathcal{L}(f) = \int_0^{\infty} e^{-st} f(t)\, \mathrm{d}t$ which is a function of s.*

Note also that for $f(t) = e^{kt}$ where $k \in \mathbb{R}$,

$$
\int_0^{\infty} e^{-st} e^{kt}\, \mathrm{d}t = \int_0^{\infty} e^{(k-s)t}\, \mathrm{d}t = \left[\frac{e^{(k-s)t}}{k-s}\right]_0^{\infty} = \frac{1}{s-k}
$$

provided that $s > k$.

Negative Binomial Distribution As A Mixed Poisson Distribution

The mixed Poisson distribution and the negative binomial distribution are both over dispersed. In fact the negative binomial distribution may be expressed in terms of a mixed Poisson distribution. In the probability generating function of then negative binomial put $s = 1 - z$ to give

$$
\begin{aligned}
\Pi(z) &= \left(\frac{p}{1 - z(1-p)}\right)^n \\
&= \left(\frac{p}{1 - (1-s)(1-p)}\right)^n \\
&= \left(\frac{\frac{p}{1-p}}{\frac{p}{1-p} + s}\right)^n
\end{aligned}
$$

But this is now in the form of the Laplace transform of an exponential function, so putting $\beta = \frac{p}{1-p}$ gives

$$
\begin{aligned}
&= \int_0^{\infty} \frac{e^{-sx}\beta^n x^{n-1} e^{-\beta x}}{(n-1)!}\, \mathrm{d}x \\
&= \int_0^{\infty} \frac{e^{-y}\beta^n y^{n-1}}{(s+\beta)^n (n-1)!}\, \mathrm{d}y \quad \text{where } y = (s+\beta)x \\
&= \frac{\beta^n}{(s+\beta)^n} \quad \text{since } \int_0^{\infty} e^{-t} t^n\, \mathrm{d}t = n!
\end{aligned}
$$

Hence the random variable Λ from which the Poisson parameter is observed has

$$\Pr\{\Lambda = \lambda\} = \frac{\beta^n \lambda^{n-1} e^{-\beta\lambda}}{(n-1)!}$$

which is a gamma distribution.

Mixed Binomial Distribution

Allow the binomial parameter p to vary in the interval $(0, 1)$ according to some distribution with probability density function $f(p)$. Then for the binomial random variable X,

$$\Pr\{X = x\} = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} f(p)\,\mathrm{d}p$$

One particular choice is the beta distribution, which gives

$$f(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)}$$

where $\alpha, \beta \in \mathbb{R}^+$ and B is the beta function,

$$B(\alpha, \beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1}\,\mathrm{d}p = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

note that

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

and that if $x \in \mathbb{N}$ then $\Gamma(x) = (x-1)!$. Hence the probability function may be calculated as follows.

$$\begin{aligned}
p_x &= \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)}\,\mathrm{d}p \\
&= \frac{\binom{n}{x}}{B(\alpha, \beta)} \int_0^1 p^{x+\alpha-1}(1-p)^{n-x+\beta-1}\,\mathrm{d}p \\
&= \binom{n}{x} \frac{B(x+\alpha, n-x+\beta)}{B(\alpha, \beta)} \\
&= \binom{n}{x} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(n+\alpha+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \qquad (42)
\end{aligned}$$

Now, $\Gamma(x+1) = x\Gamma(x)$ and hence

$$\begin{aligned}
\Gamma(x+\alpha) &= (x+\alpha-1)(x+\alpha-2)\ldots\alpha\Gamma(\alpha) \\
\text{so } \frac{\Gamma(x+\alpha)}{\Gamma(\alpha)} &= \prod_{i=0}^{x-1} (\alpha+i)
\end{aligned}$$

where it is taken by definition that $\prod_{i=0}^{-1} a_i = 1$ where a_i is anything. (Consistently, a sum over an invalid range is taken to be zero.) Hence returning to equation (42),

$$\begin{aligned}
p_x &= \binom{n}{x} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(n+\alpha+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \\
&= \binom{n}{x} \left(\prod_{i=0}^{x-1} \alpha+i\right) \left(\prod_{i=0}^{n-x-1} \beta+i\right) \left(\prod_{i=0}^{n-1} \beta+\alpha+i\right)^{-1}
\end{aligned}$$

Now divide by $(\alpha+\beta)^n$ to give

$$\begin{aligned}
p_x &= \binom{n}{x}\left(\prod_{i=0}^{x-1}\frac{\alpha+i}{\alpha+\beta}\right)\left(\prod_{i=0}^{n-x-1}\frac{\beta+i}{\alpha+\beta}\right)\left(\prod_{i=0}^{n-1}\frac{\beta+\alpha+i}{\alpha+\beta}\right)^{-1}\\
&= \binom{n}{x}\left(\prod_{i=0}^{x-1}\frac{\alpha}{\alpha+\beta}+i\theta\right)\left(\prod_{i=0}^{n-x-1}1-\frac{\beta}{\alpha+\beta}+i\theta\right)\left(\prod_{i=0}^{n-1}1+i\theta\right)^{-1} \quad \text{where } \theta=\frac{1}{\alpha+\beta}
\end{aligned}$$

If $\theta = 0$ then this expression reduces to the binomial distribution with $p = \frac{\alpha}{\alpha+\beta}$.

Something about a natural re-parameterisation where $\alpha \mapsto \frac{\alpha}{\alpha+\beta}$ and $\beta \mapsto \frac{1}{\alpha+\beta}$.

To find the mean and variance of this "beta-binomial" distribution, the probability generating function is calculated so that Section 34.3.3 and Section 34.3.3 can be applied.

$$\begin{aligned}
\pi(z) &= \sum_{x=0}^{n} z^x p_x\\
&= \sum_{x=0}^{n} z^x \int_0^1 \binom{n}{x} p^x(1-p)^{n-x}\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}\,\mathrm{d}p\\
&= \int_0^1 \left(\binom{n}{x}(pz)^x(1-p)^{n-x}\right)\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}\,\mathrm{d}p\\
&= \int_0^1 (1-p+pz)^n\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}\,\mathrm{d}p\\
&= \mathbb{E}\left(1-P+zP\right)
\end{aligned}$$

where the expectation is taken over the beta distributed random variable P. To find the mean,

$$\mu = \left.\frac{\mathrm{d}\pi}{\mathrm{d}z}\right|_{z=1}$$

$$\left.\int_0^1 np(1-p+pz)^{n-1}\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}\,\mathrm{d}p\right|_{z=1}$$

$$\begin{aligned}
&= n\int_0^1 \frac{p^{\alpha}(1-p)^{\beta-1}}{B(\alpha,\beta)}\,\mathrm{d}p\\
&= \frac{nB(\alpha+1,\beta)}{B(\alpha,\beta)}\\
&= \frac{n\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\\
&= \frac{n\alpha}{\alpha+\beta}
\end{aligned}$$

which is consistent with the interpretation of p_x being a binomial probability with $p = \frac{\alpha}{\alpha+\beta}$. Now for the variance,

$$\begin{aligned}
\left.\frac{\mathrm{d}^2\pi}{\mathrm{d}z^2}\right|_{z=1} &= \left.\int_0^1 n(n-1)p^2(1-p+pz)^{n-2}\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}\,\mathrm{d}p\right|_{z=1} \\
&= n(n-1)\int_0^1 \frac{p^{\alpha+1}(1-p)^{\beta-1}}{B(\alpha,\beta)}\,\mathrm{d}p \\
&= \frac{B(\alpha+2,\beta)}{B(\alpha,\beta)} \\
&= \frac{n(n-1)\Gamma(\alpha+2)\Gamma(\beta)}{\Gamma(\alpha+\beta+2)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \\
&= \frac{n(n-1)\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}
\end{aligned}$$

Hence using Section 34.3.3,

$$\begin{aligned}
\sigma^2 &= \frac{n(n-1)\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} + \frac{n\alpha}{\alpha+\beta} - \left(\frac{n\alpha}{\alpha+\beta}\right) \\
&= \vdots \\
&= \mu\left(1-\frac{\mu}{n}\right)\left(1+\frac{(n-1)\theta}{1+\theta}\right)
\end{aligned}$$

(34.3.5) Alternative Constructions Of Discrete Distributions

General Model

A Markov process is a random variable $X(t)$ for "time" t for which $X(0) = 0$ and takes integer values such that

$$\begin{aligned}
\Pr\{X(t+\delta t) = i+1 \mid X(t) = i\} &= \lambda_i \delta t \\
\Pr\{X(t+\delta t) = i \mid X(t) = i\} &= 1 - \lambda_i \delta t
\end{aligned}$$

The unconditional probabilities $p_i(t) = \Pr\{X(t) = i\}$ are of interest, and it will be shown that particular choices for the sequence $\{\lambda_i\}_{i=0}^{\infty}$ yield probabilities $p_i(t)$ according to familiar distributions.

$$\begin{aligned}
\Pr\{X(t+\delta t) = j\} &= \sum_{i=0}^{\infty} \Pr\{X(t+\delta t) = j \mid X(t) = i\}\Pr\{X(t) = i\} \\
&= \sum_{i\in\{j,j+1\}} \Pr\{X(t+\delta t) = j \mid X(t) = i\}\Pr\{X(t) = i\} \\
p_j(t+\delta t) &= (1-\lambda_j\delta t)p_j(t) + (\lambda_{j-1}\delta t)p_{j-1}(t) \\
\frac{p_j(t+\delta t) - p_j(t)}{\delta t} &= -\lambda_j p_j(t) + \lambda_{j-1}p_{j-1}(t) \\
\frac{\mathrm{d}p_j(t)}{\mathrm{d}t} &= -\lambda_j p_j(t) + \lambda_{j-1}p_{j-1}(t)
\end{aligned}$$

though of course this can only hold for $j \geqslant 1$, which begs the question as to what happens for $j = 0$. In this case $p_0(t+\delta t) = (1-\lambda_0\delta t$ and so

$$\frac{\mathrm{d}p_0(t)}{\mathrm{d}t} = -\lambda_0 p_0(t)$$

Hence the following system of differential difference equations has been determined.

$$\frac{\mathrm{d}p_j(t)}{\mathrm{d}t} = -\lambda_j p_j(t) + \lambda_{j-1}p_{j-1}(t) \quad \frac{\mathrm{d}p_0(t)}{\mathrm{d}t} = -\lambda_0 p_0(t) \quad p_0(0) = 1 \quad p_j(0) = 1\ \forall j \geqslant 1 \tag{43}$$

Clearly $p_0(t) = e^{-\lambda_0 t}$. For $j \geqslant 1$,

$$\begin{aligned}
\frac{\mathrm{d}p_j(t)}{\mathrm{d}t} &= -\lambda_j p_j(t) + \lambda_{j-1}p_{j-1}(t) \\
\frac{\mathrm{d}p_j(t)}{\mathrm{d}t} + \lambda_j p_j(t) &= \lambda_{j-1}p_{j-1}(t) \\
\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{\lambda_j t}p_j(t)\right) &= e^{-\lambda_j t}\lambda_{j-1}p_{j-1}(t) \\
p_j(t) &= e^{\lambda_j t}\int_0^t e^{-\lambda_j s}\lambda_{j-1}p_{j-1}(t)\ \mathrm{d}s \\
&= \int_0^t e^{-\lambda_j(s-t)}\lambda_{j-1}p_{j-1}(t)\ \mathrm{d}s
\end{aligned}$$

and so for any given sequence $\{\lambda_j\}_{j=0}^{\infty}$ this generates a discrete probability distribution $p_0(t), p_1(t), \ldots$ for chosen t. It is usual to take $t = 1$ for simplicity.

In fact for any discrete probability distribution with probabilities $\pi_0, \pi_1, \ldots$ a sequence $\{\lambda_j\}_{j=0}^{\infty}$ can be found which yields the given probabilities.

A few particular choices for $\{\lambda_j\}_{j=0}^{\infty}$ are now examined more carefully.

Constant λ_j

Put $\lambda_j = \lambda$ for all j then equation (43) becomes

$$\frac{\mathrm{d}p_j(t)}{\mathrm{d}t} = -\lambda p_j(t) + \lambda p_{j-1}(t) \quad \frac{\mathrm{d}p_0(t)}{\mathrm{d}t} = -\lambda p_0(t) \quad p_0(0) = 1 \quad p_j(0) = 1\ \forall j \geqslant 1$$

Rather than try to solve these equations directly, the probability generating function provides an easier alternative.

$$\begin{aligned}
\pi(z,t) &= \sum_{j=0}^{\infty} p_j(t)z^j \\
\frac{\partial \pi}{\partial t} &= \sum_{j=0}^{\infty}\frac{\mathrm{d}p_j(t)}{\mathrm{d}t}z^j \\
&= -\sum_{j=0}^{\infty}\lambda p_j(t)z^j + \sum_{j=1}^{\infty}\lambda p_{j-1}(t)z^j \\
&= -\lambda\pi(z,t) + \lambda z\pi(z,t) \\
&= -\lambda(1-z)\pi(z,t) \\
\pi(z,t) &= e^{-\lambda(1-z)t}f(z)
\end{aligned}$$

for some arbitrary function f. But using the initial conditions $\pi(z,0) = 1$ from which clearly $f(z) = 1$ and hence

$$\pi(z,t) = e^{-\lambda(1-z)t}$$

But this is the probability generating function of the Poisson distribution, so

$$p_j(t) = \frac{e^{-\lambda t}(\lambda t)^j}{j!}$$

Linear Increasing λ_j

Put $\lambda_j = a + bj$ for $a, b \in \mathbb{R}^+$ then equation (43) becomes

$$\frac{\mathrm{d}p_j(t)}{\mathrm{d}t} = -(a+bj)p_j(t) + (a+b(j-1))p_{j-1}(t) \quad \frac{\mathrm{d}p_0(t)}{\mathrm{d}t} = -ap_0(t) \quad p_0(0) = 1 \quad p_j(0) = 1\ \forall j \geqslant 1$$

Heading again for the probability generating function,

$$\begin{aligned}
\frac{\partial \pi}{\partial t} &= \sum_{j=0}^{\infty} \frac{\mathrm{d}p_j(t)}{\mathrm{d}t} z^j \\
&= -\sum_{j=0}^{\infty}(a+bj)p_j z^j + \sum_{j=1}^{\infty}(a+b(j-1))p_{j-1}z^j \\
&= -a\sum_{j=0}^{\infty} p_j z^j - bz\sum_{j=0}^{\infty} jp_j z^{j-1} + az\sum_{j=1}^{\infty} p_{j-1}z^{j-1} bz^2 \sum_{j=1}^{\infty}(j-1)p_{j-1}z^{j-2} \\
&= -a(1-z)\pi(z,t) - \\
\frac{\partial \pi}{\partial t} + bz(1-z)\frac{\partial \pi}{\partial z} &= a(1-z)\pi
\end{aligned}$$

Now, incredibly informally, in general

$$\frac{\partial \pi}{\partial t}\,\mathrm{d}t + \frac{\partial \pi}{\partial z}\,\mathrm{d}z = \ \mathrm{d}\pi$$

and hence put $\mathrm{d}t = 1$, $\mathrm{d}z = bz(1-z)$, and $\partial \pi = -a(1-z)\pi)$ to give

$$\frac{\mathrm{d}t}{1} = \frac{\mathrm{d}z}{bz(1-z)} = \frac{\mathrm{d}\pi}{-a(1-z)\pi}$$

with another feat of alarming informality, this gives

$$\begin{aligned}
\frac{\mathrm{d}z}{\mathrm{d}t} &= bz(1-z) & \frac{\mathrm{d}\pi}{\mathrm{d}t} &= \frac{-a(1-z)\pi}{bz(1-z)} \\
\int \frac{1}{bz(1-z)}\,\mathrm{d}z &= \int 1\,\mathrm{d}t & &= \frac{-a\pi}{bz} \\
\frac{1}{b}\int \frac{1}{z} + \frac{1}{1-z}\,\mathrm{d}z &= \int 1\,\mathrm{d}t & b\int \frac{1}{\pi}\,\mathrm{d}\pi &= -a\int \frac{1}{z}\,\mathrm{d}z \\
\frac{1}{b}\left(\ln z + \ln(1-z)\right) &= t + C & b\ln \pi + a\ln z &= D \\
\frac{1}{b}\ln\left(\frac{z}{1-z}\right) &= t + C & \ln\left(\pi^b z^a\right) &= D \\
C &= \frac{ze^{-bt}}{1-z} & D &= \pi^b z^a
\end{aligned}$$

Hence if z and t vary such that $\frac{ze^{-bt}}{1-z}$ is constant then $\pi^b z^a$ is constant. Hence

$$\pi^b z^a = \phi\left(\frac{ze^{-bt}}{1-z}\right)$$

for some arbitrary function ϕ. To find ϕ observe that $\pi(z,0) = 1$ and therefore

$$z^a = \phi\left(\frac{z}{1-z}\right)$$

and by putting $x = \frac{z}{1-z}$ it is evident that

$$\phi(x) = \left(\frac{x}{1+x}\right)^a$$

and hence

$$\pi(z,t) = \left(\frac{e^{-bt}}{1-z(1-e^{-bt})}\right)^{\frac{a}{b}}$$

This is the same probability generating function as for the negative binomial with $p = e^{-bt}$ and $n = \frac{a}{b}$. Therefore

$$p_j(t) = \frac{\left(\frac{a}{b}+(j-1)\right)\left(\frac{a}{b}+(j-2)\right)\ldots\left(\frac{a}{b}\right)}{j!}\left(1-e^{-bt}\right)^j\left(e^{bt}\right)^{\frac{a}{b}}$$

Linear Decreasing λ_j

Put $\lambda_j = a(n-j)$ for $0 \leqslant j \leqslant n$ and $\lambda_j = 0$ for $j \geqslant n$ then equation (43) becomes

$$\frac{\mathrm{d}p_j(t)}{\mathrm{d}t} = -a(n-j)p_j(t) + a(n-(j-1))p_{j-1}(t) \quad \frac{\mathrm{d}p_0(t)}{\mathrm{d}t} = -ap_0(t) \quad p_0(0) = 1 \quad p_j(0) = 1\ \forall 1 \leqslant j \leqslant n$$

now calculating the probability generating function,

$$\begin{aligned}
\frac{\partial\pi}{\partial t} &= \sum_{j=0}^{n} \frac{\mathrm{d}p_j(t)}{\mathrm{d}t} z^j \\
&= -\sum_{j=0}^{n} a(n-j)p_j(t)z^j + \sum_{j=1}^{n} a(n-(j-1))p_{j-1}(t)z^j \\
&= -\sum_{j=0}^{n} a(n-j)p_j(t)z^j + \sum_{j=1}^{n+1} a(n-(j-1))p_{j-1}(t)z^j \\
&= -an\sum_{j=0}^{n} p_j(t)z^j + az\sum_{j=0}^{n} jp_j(t)z^{j-1} + anz\sum_{j=1}^{n+1} p_{j-1}(t)z^{j-1} - az\sum_{j=1}^{n+1} p_{j-1}(t)z^{j-2}(j-1) \\
&= -an\pi + az\pi + az\frac{\partial\pi}{\partial z} + an\pi - az^2\frac{\partial\pi}{\partial z} \\
\frac{\partial\pi}{\partial t} - az(1-z)\frac{\partial\pi}{\partial z} &= -an(1-z)\pi
\end{aligned}$$

This partial differential equation has similar form to that obtained when λ_j was a linear decreasing sequence. This equation may be obtained by replacing b with $-a$ and $-a$ with $-an$. Hence

$$\begin{aligned}
\pi &= \left(\frac{e^{at}}{1-z(1-e^{at})}\right)^{\frac{an}{-a}} \\
&= \left(e^{-at} + z(1-e^{-at}\right)^n
\end{aligned}$$

which is the probability generating function for the binomial distribution, and hence

$$p_j(t) = \binom{n}{j}(1-e^{-at})^j(e^{-at})^{n-j}$$

for $0 \leqslant j \leqslant n$.