Binomial and Poisson Models

In this lecture we introduce two more statistical models which are based on two discrete probability distributions:

- The binomial model
- The poisson model

The distributions can be used to model the number of deaths observed in a mortality investigation.

The aim of the models is to derive estimates of the true values of q_x using the binomial model or μ_x using the Poisson model.

The Binomial Model

Observe N identical, independent lives aged x exactly for one year, and record the number d who die. Then d is a sample value of a random variable D

If we suppose that each life dies with probability q_x and survives with probability $1 - q_x$, then D has a binomial distribution with parameters N and q_x

The death or survival of each life can be represented by an independent Bernoulli trial with associated probabilities of q_x and $1 - q_x$ respectively.

Here q_x refers to the *initial* rate of mortality.

For the thought out experiment outlined above, the probability that exactly d deaths will occur during the year is

$$P[D = d] = \binom{N}{d} q^d (1-q)^{N-d}$$

Proof

Since we have assumed that deaths operate independently, the probability that a specified d individuals will die during the year and the remaining $N - d$ will not is $q^d(1-q)^{N-d}$.

However we need to multiply this probability by the combinatorial factor $\binom{N}{d} = \frac{N!}{d!(N-d)!}$ which is the number of ways the d deaths would be chosen.

Estimating q_x from the data

The intuitive estimate of q_x is $\hat{q}_x = \frac{d}{N}$ and this is also the maximum likelihood estimate.

proof

Under the binomial model, the likelihood of recording exactly d deaths if the rate of mortality is q is

$$L(q) = \binom{N}{q} q^d (1-q)^{N-d}$$

which can be maximised by maximising its log

$$\log L(q) = \log \binom{N}{d} + d \log q + (N - d) \log(1 - q)$$

Differentiating wrt to q

$$\frac{\partial}{\partial q} \log L(q) = \frac{d}{q} - \frac{N-d}{1-q}$$

This is zero at the value $\hat{q}$ such that:

$$d(1-\hat{q}) = (N-d)\hat{q} \Longrightarrow \hat{q} = \frac{d}{N}$$

This is maximum since

$$\frac{\partial^2}{\partial q^2}\log L(q) = -\frac{d}{q^2} - \frac{N-d}{(1-q)^2} < 0$$

The maximum likelihood estimate is the observed value of the corresponding maximum likelihood estimator

$$\tilde{q}_x = \frac{D}{N}$$

The corresponding estimator $\tilde{q}_x$ has:

- Mean q_x it is unbiased
- Variance $\frac{q_x(1-q_x)}{N}$

The observed number of deaths D has a Binomial (N, q_x) distribution which mean that it has a mean Nq_x and variance $Nq_x(1-q_x)$ so

$$E(\tilde{q}_x) = E(\frac{D}{N}) = \frac{Nq_x}{N} = q_x$$

The asymptotic distribution of $E(\tilde{q}_x)$ is

$$E(\tilde{q}_x) \sim N\Big(q_x, \frac{q_x(1-q_x)}{N}\Big)$$

This is the binomial model of mortality.

The Poisson Model

The Poisson distribution is used to model the number of 'rare' events occurring during some period of time.

A random variable X is said to have a Poisson distribution with mean λ if the probability function of X is

$$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Let E_x^c denote the total observed waiting time.

If we assume that we observe N individuals as before and that the force of mortality is a constant μ, then a Poisson model is given by the assumption that D has a Poisson distribution with parameter μE_x^c That is

$$P[D=d] = \frac{e^{-\mu}E_x^c(\mu E_x^c)^d}{d!}$$

The Poisson likelihood leads to the following estimator of μ

$$\tilde{\mu} = \frac{D}{E_x^c}$$

proof

The likelihood of observing d deaths if the true value of the hazard rate is μ is

$$L(\mu) = \frac{(\mu E_x^c)^d e^{-\mu E_x^c}}{d!}$$

which can be maximised by maximising its log

$$\log L(\mu) = d(\log \mu + \log E_x^c) - \mu E_x^c - \log d!$$

Differentiating w.r.t μ

$$\frac{\partial}{\partial \mu} \log L(\mu) = \frac{d}{\mu} - E_x^c$$

which is zero when

$$\hat{\mu} = \frac{d}{E_x^c}$$

The estimator $\tilde{\mu}$ has the following properties

- $E[\tilde{\mu}] = \mu$
- $Var[\tilde{\mu}] = \frac{\mu}{E_x^c}$

The asymptotic distribution of $\tilde{\mu}$ is

$$\tilde{\mu} \sim N\Big(\mu, \frac{\mu}{E_x^c}\Big)$$

Exposed to Risk

Central exposed to risk is the total waiting time which features in both two-state markov model and the poisson model.

The central exposed to risk is a natural quantity intrinsically observable even if the observation may be incomplete in practice.

Homogeneity

The Poisson models are based on the assumption that we can observe groups of identical lives or homogeneous groups.

A group of lives with different characteristics is said to be *heterogenous*

As a result of this heterogeneity, our estimate of the mortality rate would be the estimate of the average rate over the whole group of lives.

Example

consider a country in which 50% of the population are smokers. If $\mu_{40} = 0.001$ for non-smokers and $\mu_{40} = 0.002$ for smokers, then a mortality investigation based on the entire population may lead us to the estimate $\hat{\mu}_{40} = 0.0015$ An insurance company that calculates its premiums using this average figure would overcharge non-smokers and undercharge smokers.

The solution is subdivide our data according to characteristics known, from experience, to have a significant effect on mortality. This ought to reduce the heterogeneity of each class.