4 Testing Hypotheses

The next lectures will look at tests, some in an actuarial setting, and in the last subsection we will also consider tests applied to graduation.

4.1 Tests in the regression setting

1) A package will produce a test of whether or not a regression coefficient is 0. It uses properties of mle's. Let the coefficient of interest be b say. Then the null hypothesis is $\mathrm{H_O} : b = 0$ and the alternative is $\mathrm{H}_A : b \neq 0$. At the 5% significance level, $\mathrm{H_O}$ will be accepted if the *p-value* $p > 0.05$, and rejected otherwise.

2) In an AL parametric model if α is the shape parameter then we can test $\mathrm{H_O} : \log \alpha = 0$ against the alternative $\mathrm{H}_A : \log \alpha \neq 0$. Again mle properties are used and a p-value is produced as above. In the case of the Weibull if we accept $\log \alpha = 0$ then we have the simpler exponential distribution (with $\alpha = 1$).

3) We have already mentioned that, to test Weibull v. exponential with null hypothesis $\mathrm{H_O}$: exponential is an acceptable fit, we can use

$$2 \log \widehat{L}_{weib} - 2 \log \widehat{L}_{\exp} \sim \chi^2(1), \text{ asymptotically.}$$

4.2 Non-parametric testing of survival between groups

We will just consider the case where the data splits into two groups. There is a relatively easy extension to k(>2) groups.

We need to set up some notation.

Notation: $\begin{cases} d_{i1} = \#\text{ events at } t_i \text{ in group 1,} \\ n_{i1} = \#\text{ in risk set at } t_i \text{ from group 1,} \end{cases}$

similarly d_{i2}, n_{i2}.

Event times are $(0 <) t_1 < t_2 < \cdots < t_m$.

$\begin{cases} d_i = \#\text{ events at } t_i \\ n_i = \#\text{ in risk set at } t_i \end{cases}$

$d_i = d_{i1} + d_{i2}, n_i = n_{i1} + n_{i2}$

All the tests are based on the following concepts:

Observed # events in group 1 at time $t_i, = d_{i1}$

Expected # events in group 1 at time $t_i, = n_{i1} \frac{d_i}{n_i}$ under the null hypothesis below

H_{O} : there is no difference between the hazard rates of the two groups

Usually the alternative hypothesis for all tests is 2-sided and simply states that there is a difference in the hazard rates.

Finally using a hypergeometric distribution under the assumption that n_{i1}, d_i, n_i are each fixed, the computed variance of the number of events at time t_i in group 1, is

$$\frac{n_{i1} n_{i2} (n_i - d_i) d_i}{n_i^2 (n_i - 1)} .$$

The **test statistics** used take the form

$$Z = \frac{\sum_1^m W(t_i)\left(d_{i1} - n_{i1}\frac{d_i}{n_i}\right)}{\sqrt{\sum_1^m W(t_i)^2 \frac{n_{i1}n_{i2}(n_i-d_i)d_i}{n_i^2(n_i-1)}}}$$

where $W(t_i)$ are weights, and Z has an approximate standard normal distribution when H_O is true. Clearly $Z^2 \sim \chi^2(1)$, also under these conditions. Note that approximate normality works because the hazards at different times are asymptotically independent.

The tests

(1) $W(t_i) = 1, \forall i$. This is the **log rank test**, and is the test in most common use. A criticism is that it can give too much weight to the later event times when numbers in the risk sets may be relatively small. The log rank test is aimed at detecting a consistent difference between hazards in the two groups and is best placed to consider this alternative when the proportional hazard assumption applies.

(2) Peto's test uses a weight dependent on a modified estimated survival function, estimated for the whole study. The modified estimator is

$$\widetilde{S}(t) = \prod_{t_i \leq t} \frac{n_i + 1 - d_i}{n_i + 1}$$

and the suggested weight is then

$$W(t_i) = \widetilde{S}(t_{i-1})\frac{n_i}{n_i + 1}$$

This has the advantage of giving more weight to the early events and less to the later ones where the population remaining is smaller.

(3) $W(t_i) = n_i$ has also been suggested (Gehan, Breslow). This again downgrades the effct of the later times.

(4) Fleming and Harrington suggested a range of tests using

$$W(t_i) = \left(\widehat{S}(t_{i-1})\right)^p \left(1 - \widehat{S}(t_{i-1})\right)^q$$

where $\widehat{S}$ is the Kaplan-Meier survival function, estimated for all the data. Then $p = q = 0$ gives the logrank test and $p = 1, q = 0$ gives a test very close to Peto's test and is called the Fleming-Harrington test. If we were to set $p = 0, q > 0$ this would emphasise the later event times if needed for some reason.

Worries with these are:

a) undue unfluence of censoring if pattern of censoring is very different in one group

b) over-emphasis of tail if small numbers remain (logrank test).

All of the above can be extended to more than 2 groups. They test for a consistent (same sign) difference between hazards.

Unusually we may be interested in "crossing" of hazards (possibly also survival functions). There may be some interaction going on between group and survival time and so a non-PH effect of group. Then different tests have to be applied. For instance it is clear that we could use the standard χ^2 test

$$\sum \frac{(O-E)^2}{E}$$

5 Testing in an Actuarial Situation

5.1 Notation

First some notation: compare the notation in survival analysis in a general setting and in the population setting used by actuaries.

Future lifetime of a newborn is a random variable T distributed on $[0, \varpi]$, where ϖ represents the maximum age (usually chosen to lie in the range 100-120)

The future lifetime after age x is denoted by T_x.

$$\begin{aligned}
F_T(t) &= \Pr(T \leq t) \\
S_T(t) &= \overline{F}_T(t) = 1 - F_T(t) = \Pr(T > t) \\
F_x(t) &= \Pr(T_x \leq t) \\
&= \Pr(T \leq x + t | T > x) \\
&= \frac{\Pr(x < T \leq x + t)}{S_T(x)} \\
&= \frac{F_T(x+t) - F_T(x)}{S_T(x)}
\end{aligned}$$

Then associated with this, the probability of death by t given alive at x is

$${}_tq_x = F_x(t)$$

and the probability of being alive at t given alive at x is

$${}_tp_x = 1 - {}_t\, q_x = S_x(t).$$

We then have to consider two possible distinctions in models used. That is both *discrete* and *continuous* models are in use. We recapitulate some of the earlier lecture notes here.

Definition (rate of mortality)

In a **discrete** model the probability of death within one year of birthday age x is called the **rate of mortality** and is $q_x = {}_1\, q_x$.

Definition (the force of mortality)

In a **continuous** model the hazard rate

$$\mu_x = \lim_{h \downarrow 0} \frac{\Pr(T \leq x + h | T > x)}{h} \quad (= \text{hazard rate})$$

is called the **force of mortality.**

Note that μ_{x+t} can be used for fixed age x and $0 \leq t < \varpi - x$. Given T_x in the continuous model, it has **probability density function**

$$f_x(t) = \mu_{x+t} S_x(t) = \mu_{x+t} \left(1 - F_x(t)\right).$$

The **complete expectation of life** after age x is

$$\begin{aligned} \mathrm{E}(T_x) &= \int_0^{\varpi - x} t f_x(t) \mathrm{dt} \\ &= \int_0^{\varpi - x} S_x(t) \mathrm{dt} \end{aligned}$$

The random variable representing the **curtate future lifetime** is $K_x = [T_x]$, and so T_x rounded down to the nearest integer. The curtate expectation of life is just $\mathrm{E}(K_x)$.

The interest lies in mortality and predictions of death if in life insurance and survival if in pensions. Whichever setting there is a choice of discrete or continuous model.

5.2 The binomial model

Ideally we observe n identically distributed, independent lives aged x for exactly 1 year. we record the number d_x who die. Using the notation set up for the discrete model, a life dies with probability q_x within the year.

Hence D_x, the random variable representing the numbers dying in the year conditional on n alive at the beginning of the year, has distribution

$$D_x \sim B(n, q_x)$$

giving a maximum likelihood estimator

$$\widehat{q}_x = \frac{D_x}{n}, \text{ with } \mathrm{var}\widehat{q}_x = \frac{q_x \left(1 - q_x\right))}{n}$$

where using previous notation we have set $l_x = n$.

The trouble is of course that with real data we may observe the ith life in an interval (a_i, b_i), $0 < a_i < b_i < 1$. In this case

$$\Pr(D_{xi} = 1) =_{b_i - a_i} q_{x + a_i}$$

Hence

$$\mathrm{E}D_x = \mathrm{E}\left(\sum D_{xi}\right) = \sum_{i=1}^{n} {}_{b_i-a_i}q_{x+a_i}$$

To evaluate we have to make an assumption about the rate of mortality over a period of less than 1 year.

There are various possibilities of which common ones are (for $0 < t < 1$)

(i) uniform on (0,1) if death occurs in (0,1), giving $F_x(t) = tq_x$

(ii) Balducci: ${}_{1-t}q_{x+t} = (1-t)q_x$, so that the probability of failing in the remaining time $1-t$, having survived to $x+t$, is the product of the time left and the probability of failure in $(x, x+1)$

(iii) there is a constant force of mortality over the year so that ${}_tq_x = 1 - \mathrm{e}^{-\mu_{x+\frac{1}{2}}t}$, where $\mu_{x+\frac{1}{2}} = \mu_{x+t}$ for $0 < t < 1$.

If Balducci assumptions are used then it can be shown that a sensible estimator is the actuarial estimator, with observed value

$$\widetilde{q}_x = \frac{d_x}{E_x^c + \frac{1}{2}d_x}$$

The denominator, $E_x^c + \frac{1}{2}d_x$, comprises the observed time at risk (also called central exposed to risk) within the interval $(x, x+1)$, added to 1/2 the number of deaths (assumes deaths evenly spread over the interval). This is an estimator for E_x which is the inital exposed to risk and is what is required for the binomial model.

NB assumptions (i)-(iii) collapse to the same model, essentially (i), if $\mu_{x+\frac{1}{2}}$ is very small, since all become ${}_tq_x \approx t\mu_{x+\frac{1}{2}}$, $0 < t < 1$.

Definitions, within year $(x, x+1)$

a) E_x^c = observed total time (in years) at risk = **central exposed to risk**

b) $E_x^0 (= E_x)$ = initial exposed to risk = # in risk set at age x, with approximation $E_x \approx E_x^c + \frac{1}{2}d_x$, if required.

5.3 The Poisson model

Under the assumption of a constant hazard rate (force of mortality) $\mu_{x+\frac{1}{2}}$ over the year $(x, x+1]$, with observed years at risk E_x^c, then if D_x represents the numbers dying in the year the model uses

$$\Pr(D_x = k) = \frac{\left(\mu_{x+\frac{1}{2}}E_x^c\right)^k \mathrm{e}^{-\mu_{x+\frac{1}{2}}E_x^c}}{k!}, \quad k = 0, 1, 2, \cdots$$

which is an approximation to the 2-state model, and which in fact yields the same likelihood.

The estimator for the constant force of mortality over the year is

$$\widetilde{\mu}_{x+\frac{1}{2}} = \frac{D_x}{E_x^c}, \quad \text{with estimate } \frac{d_x}{E_x^c} \ .$$

Under the Poisson model we therefore have that

$$\mathrm{var}\widetilde{\mu}_{x+\frac{1}{2}} = \frac{\mu_{x+\frac{1}{2}} E_x^c}{\left(E_x^c\right)^2} = \frac{\mu_{x+\frac{1}{2}}}{E_x^c} \ .$$

So the estimate will be

$$\mathrm{var}\widetilde{\mu}_{x+\frac{1}{2}} \approx \frac{d_x}{\left(E_x^c\right)^2} \ .$$

If we compare with the **2-state stochastic model** over year $(x, x+1)$, assuming constant $\mu = \mu_{x+\frac{1}{2}}$, then the likelihood is

$$L = \prod_1^n \mu^{\delta_i} \mathrm{e}^{-\mu t_i} \ ,$$

where $\delta_i = 1$ if life i dies and $t_i = b_i - a_i$ in previous terminology (see the binomial model). Hence

$$L = \mu^{d_x} \mathrm{e}^{-\mu E_x^c}$$

and so

$$\widehat{\mu} = \frac{D_x}{E_x^c} \ .$$

The estimator is exactly the same as for the Poisson model except that both D_x and E_x^c are random variables. Using asymptotic likelihood theory we see that the estimate for the variance is

$$\mathrm{var}\widehat{\mu} \approx \frac{\mu^2}{d_x} \approx \frac{d_x}{\left(E_x^c\right)^2} \ .$$

5.4 Testing hypotheses for q_x and $\mu_{x+\frac{1}{2}}$

We note the following normal approximations:

(i) Binomial model:

$$D_x \sim B(E_x, q_x) \implies D_x \sim N(E_x q_x, E_x q_x \left(1 - q_x\right)$$

and

$$\widehat{q}_x = \frac{D_x}{E_x} \sim N\left(q_x, \frac{q_x \left(1 - q_x\right)}{E_x}\right) \ .$$

(ii) Poisson model or 2-state model:

$$D_x \sim N(E_x^c \mu_{x+\frac{1}{2}}, E_x^c \mu_{x+\frac{1}{2}})$$

and

$$\widehat{\mu}_{x+\frac{1}{2}} \sim N\left(\mu_{x+\frac{1}{2}}, \frac{\mu_{x+\frac{1}{2}}}{E_x^c}\right) \ .$$

Tests are often done using comparisons with a published **standard life table.** These can be from

a) national tables for England and Wales published every 10 years,

b) insurance company data collected by the Continuous Mortality Investigation Bureau.

A superscript "s" denotes "from a standard table", such as q_x^s and $\mu^s_{x+\frac{1}{2}}$.

Test statistics are generally obtained from the following:

Binomial:

$$z_x = \frac{d_x - E_x q_x^s}{\sqrt{E_x q_x^s \left(1 - q_x^s\right)}} \quad \left(\approx \frac{O-E}{\sqrt{V}}\right)$$

Poisson/2-state:

$$z_x = \frac{d_x - E_x^c \mu^s_{x+\frac{1}{2}}}{\sqrt{E_x^c \mu^s_{x+\frac{1}{2}}}} \quad \left(\approx \frac{O-E}{\sqrt{V}}\right) .$$

Both of these are denoted as z_x since under a null hypothesis that the standard table is adequate $Z_x \sim N(0,1)$ approximately.

5.4.1 The tests

(A) χ^2 **test**.

We take

$$X = \sum_{\text{all ages } x} z_x^2$$

This gives the sum of squares of standard normal random variables under the null hypothesis and so is a sum of $\chi^2(1)$. Therefore

$$X \sim \chi^2(m) \text{ , if } m = \#\text{ years of study.}$$

H_0 : there is no difference between the standard table and the data,

H_A : they are not the same.

It is normal to use 5% significance and so the test fails if $X > \chi^2(m)_{0.95}$.

It tests large deviations from the standard table.

Disadvantages:

1. There may be a few large deviations offset by substantial agreement over part of the table. The test will not pick this up.

2. There might be bias, that is, although not necessarily large, all the deviations may be of the same sign.

3. There could be significant groups of consecutive deviations of the same sign, even if not overall.

(B) **Standardised deviations test.**

This tries to address point 1 above. Noting that each z_x is an observation from a standard normal distribution under H_0, the real line is divided into intervals, say 6 with dividing points at $-2, -1, 0, 1, 2$. The number of z_x in each

interval is counted and compared with the expected number from a standard normal distribution. The test statistic is then

$$X = \sum_{\text{intervals}} \frac{(O-E)^2}{E} \sim \chi^2(5)$$

under the null hypothesis since this is Pearson's statistic. The problem here is that m is unlikely to be large enough to give approximate validity to the chi-square distribution. So this test is rarely appropriate.

(C) **Signs test.**

Test statistic X is given by

$$X = \#\{z_x > 0\}$$

Under the null hypothesis $X \sim B(m, \frac{1}{2})$, since the probability of a positive sign should be 1/2. This should be administered as a two-tailed test. It is underpowered since it ignores the size of the deviations but it will pick up small deviations of consistent sign, positive or negative, and so it addresses point 2 above.

(D) **Cumulative deviations test.**

This again addresses point 2 and essentially looks very similar to the logrank test between two survival curves. If instead of squaring $d_x - E_x q_x^s$ or $d_x - E_x^c \mu_{x+\frac{1}{2}}^s$, we simply sum then

$$\frac{\sum (d_x - E_x q_x^s)}{\sqrt{\sum E_x q_x^s (1-q_x^s)}} \sim N(0,1), \quad \text{approximately}$$

and

$$\frac{\sum \left(d_x - E_x^c \mu_{x+\frac{1}{2}}^s\right)}{\sqrt{\sum E_x^c \mu_{x+\frac{1}{2}}^s}} \sim N(0,1) \quad \text{approximately.}$$

H_0 : there is no bias
H_A : there is a bias.

This test addresses point 2 again, which is that the chi-square test does not test for consistent bias.

(E) There are tests to deal with consecutive bias/runs of same sign. These are called the groups of signs test and the serial correlations test. Again a very large number of years, m, are required to render these tests useful.

5.4.2 Graduation

Graduation is exactly what we would normally mean by "smoothing". Suppose that a company has collected its own data, producing estimates for either q_x or $\mu_{x+\frac{1}{2}}$. The estimates may be rather irregular from year to year and this could be an artefact of the population the company happens to have in a particular scheme. The underlying model should probably (but not necessarily) be

smoother than the raw estimates. If it is to be considered for future predictions, then **smoothing** should be considered. This is called **graduation.**

Possible methods of graduation

(A) **Parametric**

Fit a formula to the data. Possible examples are

$$\begin{aligned} \mu_x &= Bc^x \\ \mu_x &= A + Bc^x \end{aligned}$$

The first of these can be a good model for a population of middle.older age groups. It is an exponential model ($c^x = \mathrm{e}^{x \log c}$). The second has an extra additive constant which is sometimes used to model accidental deaths, regardless of age. We coud use more complicated formulae putting in polynomials in x.

(B) **Reference to a standard table**

Here q_x^0, μ_x^0 represent the graduated estimates. We could have a linear dependence

$$q_x^0 = a + bq_x^s, \quad \mu_x^0 = a + b\mu_x^s$$

or possibly a translation of years

$$q_x^0 = q_{x+k}^s, \quad \mu_x^0 = \mu_{x+k}^s$$

In general there will be some assigned functional dependence of the graduated estimate on the standard table value.

In both of these we need to get the best fit by some means.

Methods for fitting:

(i) In any of the models (binomial, Poisson, 2-state) set (say) $q_x = a + bq_x^s$ in the likelihood and use maximum likelihood estimators for the unknown parameters a, b and similarly for μ_x and other functional relationships with the standard values.

(ii) Use weighted least squares and minimise

$$\sum_{\text{all ages } x} w_x \left(\widehat{q}_x - q_x^0\right)^2 \quad \text{or}$$

$$\sum_{\text{all ages } x} w_x \left(\widehat{\mu}_{x+\frac{1}{2}} - \mu_{x+\frac{1}{2}}^0\right)^2$$

as appropriate. For the weights suitable choices are either E_x or E_x^c respectively. Alternatively we can use 1/var, where the variance is estimated for $\widehat{q}_x$ or $\widehat{\mu}_{x+\frac{1}{2}}$, respectively.

The test hypotheses we have already covered above can be used to test the graduation fit to the data, replacing $q_x^s, \mu_{x+\frac{1}{2}}^s$ by the graduated estimates. **Note that in the χ^2 test we must reduce the degrees of freedom of the χ^2 distribution by the number of parameters estimated in the model for the graduation.** For example if $q_x^0 = a + bq_x^s$, then we reduce the degrees of freedom by 2 as the parameters a, b are estimated.

Finally an alternative method of graduation is to smooth using a smoothing programme from a package. For example in Matlab the methods available could include kernel smoothing, orthogonal polynomials, cubic splines and so on. These are beyond the scope of this course.