Example 3.6: Mating of Drosophila flies

30 female virgin flies and 40 male virgin flies are put in a plastic bowl ("pornoscope") and times (in seconds) on initiatings of matings are recorded.

Two experiments: one experiment with "ebony" flies (experiment 1) and one with "oregon" flies (experiment 2)

Ebony	143	180	184	303	380	431	455	475	500	514
	521	552	558	606	650	667	683	782	799	849
	901	995	1131	1216	1591	1702	2212			
Oregon	555	742	746	795	934	967	982	1043	1055	1067
	1081	1296	1353	1361	1462	1731	1985	2051	2292	2335
	2514	2570	2970							

Let $N_h(t)$ count the number of matings in $[0, t]$ in experiment h (h=1 2)

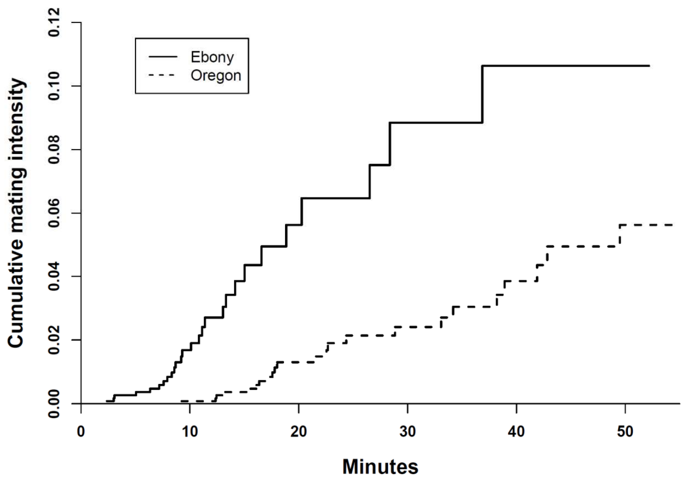
Assuming random mating, the intensity processes takes the multiplicative form

$$\lambda_h(t) = \underbrace{f_h(t)m_h(t)}_{Y_h(t)}\alpha_h(t)$$

Here $f_h(t)$ and $m_h(t)$ are the number of virgin female and male flies just before time t in experiment h and $\alpha_h(t)$ is the mating intensity in the experiment

Nelson-Aalen estimators of cumulative mating intensities

$$\hat{A}_h(t) = \sum_{T_{hj} \leq t} \frac{1}{Y_h(T_{hj})}$$



Is the observed difference in mating intensities significant?

Two-sample tests

Consider two counting process $N_1(t)$ and $N_2(t)$ with intensity processes of the multiplicative form

$$\lambda_h(t) = Y_h(t) \cdot \alpha_h(t) \qquad h = 1, 2$$

We want to test the null hypothesis

$$H_0: \alpha_1(t) = \alpha_2(t) \quad \text{for } 0 \leq t \leq t_0$$

Usually we will chose $t_0 = \tau$, i.e. upper limit of study

The common value of the $\alpha_h(t)$ under H_0 will be denoted $\alpha(t)$

Introduce the Nelson-Aalen estimators

$$\hat{A}_h(t) = \int_0^t \frac{J_h(u)}{Y_h(u)} dN_h(u)$$

We will consider the test statistic

$$Z_1(t_0) = \int_0^{t_0} L(t)\{d\hat{A}_1(t) - d\hat{A}_2(t)\}$$

where $L(t)$ is a non-negative predictable weight process that is zero whenever at least one of the $Y_h(t)$ are zero

The choice $L(t) = Y_1(t)Y_2(t)/Y_\bullet(t)$ with $Y_\bullet(t) = Y_1(t) + Y_2(t)$ gives the logrank test

The test statistic $Z_1(t_0)$ is useful for testing H_0 versus "non-crossing hazards" alternatives

If the null hypthesis holds true, we have

$$dN_h(t) = Y_h(t)\alpha(t)\,dt + dM_h(t) \qquad h = 1, 2$$

Then

$$Z_1(t_0) = \int_0^{t_0} \frac{L(t)}{Y_1(t)} dN_1(t) - \int_0^{t_0} \frac{L(t)}{Y_2(t)} dN_2(t)$$

$$= \int_0^{t_0} \frac{L(t)}{Y_1(t)} dM_1(t) - \int_0^{t_0} \frac{L(t)}{Y_2(t)} dM_2(t)$$

Thus $Z_1(t_0)$ is a mean zero martingale (in t_0) when the null hypothesis holds true

In particular $\mathrm{E}\{Z_1(t_0)\} = 0$ under H_0

Predictable variance process under H_0 :

$$\langle Z_1 \rangle(t_0) = \int_0^{t_0} \left(\frac{L(t)}{Y_1(t)}\right)^2 \lambda_1(t)dt + \int_0^{t_0} \left(\frac{L(t)}{Y_2(t)}\right)^2 \lambda_2(t)dt$$

$$= \int_0^{t_0} \frac{L^2(t)Y_\bullet(t)}{Y_1(t)\,Y_2(t)} \alpha(t)\,dt$$

This is estimated by

$$V_{11}(t_0) = \int_0^{t_0} \frac{L^2(t)}{Y_1(t)\,Y_2(t)} dN_\bullet(t)$$

The variance estimator is unbiased under H_0 (exercise 3.10)

The standardized test statistic

$$U(t_0) = \frac{Z_1(t_0)}{\sqrt{V_{11}(t_0)}}$$

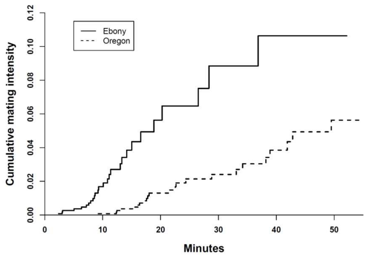
is approximately standard normal under the null hypothesis

Alternatively we may use the test statistic

$$X^2(t_0) = \frac{Z_1(t_0)^2}{V_{11}(t_0)}$$

which is approximately chi-squared with 1 df under H_0

Example 3.12: Mating of Drosophila flies



Using the logrank weights $L(t) = Y_1(t)Y_2(t)/Y_{\bullet}(t)$ we find

$$Z_1(\tau) = 15.56 \qquad V_{11}(\tau) = 8.03$$

The test statistic becomes

$$U(\tau) = \frac{15.56}{\sqrt{8.03}} = 5.49$$

which is highly significant

Table 3.2 Choice of weight process $L(t)$ for a number of two-sample tests

Test	Weight process[a]	Key references
Log-rank	$Y_1(t)Y_2(t)/Y_{\bullet}(t)$	Mantel (1966), Peto and Peto (1972)
Gehan-Breslow	$Y_1(t)Y_2(t)$	Gehan (1965), Breslow (1970)
Efron[b]	$\widehat{S}_1(t-)\widehat{S}_2(t-)J_1(t)J_2(t)$	Efron (1967)
Tarone-Ware	$Y_1(t)Y_2(t)/\sqrt{Y_{\bullet}(t)}$	Tarone and Ware (1977)
Peto-Prentice	$\widetilde{S}(t-)Y_1(t)Y_2(t)/(Y_{\bullet}(t)+1)$	Peto and Peto (1972), Prentice (1978)
Harrington-Fleming	$\widehat{S}(t-)^{\rho}Y_1(t)Y_2(t)/Y_{\bullet}(t)$	Harrington and Fleming (1982)

Some of the weight process only apply for survival data

The Harrington-Fleming test is implemented in R (default is $\rho = 0$ corresponding to the logrank test)

The test statistic and the variance estimator may be given an alternative formulation. This may be useful to obtain a better understanding of the test, and it opens for a generalization to more than two samples

We introduce the weight process

$$K(t) = \frac{L(t)Y_{\bullet}(t)}{Y_1(t)Y_2(t)}$$

The we may write:

$$Z_1(t_0) = \int_0^{t_0} K(t)\,dN_1(t) - \int_0^{t_0} K(t)\frac{Y_1(t)}{Y_{\bullet}(t)}\,dN_{\bullet}(t)$$

$$V_{11}(t_0) = \int_0^{t_0} K^2(t)\frac{Y_1(t)}{Y_{\bullet}(t)}\left(1 - \frac{Y_1(t)}{Y_{\bullet}(t)}\right)dN_{\bullet}(t)$$

Note that for the logank test we have:

$$K(t) = I\{Y_{\bullet}(t) > 0\}$$

Therefore the two-sample logrank statistic may be written

$$Z_1(t_0) = N_1(t_0) - E_1(t_0)$$

where

$$E_1(t_0) = \int_0^{t_0} \frac{Y_1(t)}{Y_{\bullet}(t)}\,dN_{\bullet}(t)$$

the "expected" number of events under the null hypothesis (exercise 3.11)

k-sample tests

We now consider k counting process $N_1(t), N_2(t), \ldots, N_k(t)$ with intensity processes of the multiplicative form

$$\lambda_h(t) = Y_h(t) \cdot \alpha_h(t) \qquad h = 1, \ldots, k$$

We want to test the null hypothesis

$$\mathrm{H}_0: \ \alpha_1(t) = \ldots.. = \alpha_k(t) \quad \text{for} \ \ 0 \le t \le t_0$$

We introduce (where δ_{hj} is a Kronecker delta)

$$Z_h(t_0) = \int_0^{t_0} K(t)\, dN_h(t) - \int_0^{t_0} K(t) \frac{Y_h(t)}{Y_\bullet(t)} dN_\bullet(t)$$

$$V_{hj}(t_0) = \int_0^{t_0} K^2(t) \frac{Y_h(t)}{Y_\bullet(t)} \left(\delta_{hj} - \frac{Y_j(t)}{Y_\bullet(t)} \right) dN_\bullet(t)$$

Note that $\sum_{h=1}^{k} Z_h(t_0) = 0$

Therefore we only consider the first $k - 1$ of the $Z_h(t_0)$'s when forming our test statistic

We introduce the $k - 1$ dimensional vector

$$\mathbf{Z}(t_0) = \left(Z_1(t_0), \ldots.., Z_{k-1}(t_0) \right)^T$$

and the $(k - 1)$x$(k - 1)$ matrix

$$\mathbf{V}(t_0) = \begin{pmatrix} V_{11}(t_0) & V_{12}(t_0) & \ldots. & V_{1,k-1}(t_0) \\ V_{21}(t_0) & V_{22}(t_0) & \ldots. & V_{2,k-1}(t_0) \\ \ldots. & \ldots. & \ldots. & \ldots. \\ V_{k-1,1}(t_0) & V_{k-1,2}(t_0) & \ldots. & V_{k-1,k-1}(t_0) \end{pmatrix}$$

Then the test statistic takes the form

$$X^2(t_0) = \mathbf{Z}(t_0)^T \mathbf{V}(t_0)^{-1} \mathbf{Z}(t_0)$$

The statistic is chi-squared distributed with the $k - 1$ df when the null hypothesis holds true

For the logrank test one may show that

$$\sum_{h=1}^{k} \frac{\left(N_h(t_0) - E_h(t_0) \right)^2}{E_h(t_0)} \le X^2(t_0) \qquad (*)$$

where $E_h(t_0) = \int_0^{t_0} \{ Y_h(t) / Y_\bullet(t) \}\, dN_\bullet(t)$

This the left-hand side of (*) provides a *conservative version* of the logrank test

Stratified tests

We now consider the situation where we have k counting process in each of m strata:

$N_{hs}(t)$ for $h = 1, \ldots, k$ and $s = 1, \ldots., m$

with intensity processes of the multiplicative form

$$\lambda_{hs}(t) = Y_{hs}(t) \cdot \alpha_{hs}(t) \qquad h = 1, \ldots., k; \ \ s = 1, \ldots, m$$

We want to test the null hypothesis

$$\mathrm{H}_0: \ \alpha_{1s}(t) = \ldots.. = \alpha_{ks}(t)$$
$$\text{for} \ \ 0 \le t \le t_0 \quad \text{for all } s = 1, \ldots, m$$

For each stratum s we define similar quantities as above:

$$Z_{hs}(t_0) = \int_0^{t_0} K_s(t)\, dN_{hs}(t) - \int_0^{t_0} K_s(t) \frac{Y_{hs}(t)}{Y_{\bullet s}(t)} dN_{\bullet s}(t)$$

$$V_{hjs}(t_0) = \int_0^{t_0} K_s^2(t) \frac{Y_{hs}(t)}{Y_{\bullet s}(t)} \left(\delta_{hj} - \frac{Y_{js}(t)}{Y_{\bullet s}(t)} \right) dN_{\bullet s}(t)$$

Further we define the $k - 1$ dimensional vectors

$$\mathbf{Z}_s(t_0) = \left(Z_{1s}(t_0), \ldots\ldots, Z_{k-1,s}(t_0) \right)^T$$

and the $(k - 1)$x$(k - 1)$ dimensional matrices

$$\mathbf{V}_s(t_0) = \left\{ V_{hjs}(t_0) \right\}_{h,j=1,\ldots,k-1}$$

We now obtain the test statistic by aggregating information over the m strata:

$$X^2(t_0) = \left(\sum_{s=1}^{m} \mathbf{Z}_s(t_0) \right)^T \left(\sum_{s=1}^{m} \mathbf{V}_s(t_0) \right)^{-1} \left(\sum_{s=1}^{m} \mathbf{Z}_s(t_0) \right)$$

The statistic is chi-squared distributed with the $k - 1$ df when the null hypothesis holds true

Regression models

Assume that we have a sample of n individuals, and let $N_i(t)$ count the observed occurrences of the event of interest for individual i as a function of (study) time t

We have the decomposition

$$\underbrace{dN_i(t)}_{\text{observation}} = \underbrace{\lambda_i(t)dt}_{\text{signal}} + \underbrace{dM_i(t)}_{\text{noise}}$$

We will consider regression models where the intensity process $\lambda_i(t)$ for individual i depends on a vector of (possibly) time-dependent covariates

$$\mathbf{x}_i(t) = (x_{i1}(t), x_{i2}(t), \ldots, x_{ip}(t))^T$$

The intensity process for individual i may be given as

$$\lambda_i(t) = \underbrace{Y_i(t)}_{\text{at risk indicator}} \cdot \underbrace{\alpha(t \mid \mathbf{x}_i)}_{\text{hazard rate (intensity)}}$$

(time-dependency of covariates suppressed in the notation)

A regression model specifies how the hazard rate depends on the covariates

We will consider two types of regression models:

- Relative risk regression models (section 4.1)
- Additive regression models (section 4.2)

Relative risk regression models

Hazard rate for individual i

$$\alpha(t \mid \mathbf{x}_i) = \underbrace{\alpha_0(t)}_{\text{baseline hazard}} \cdot \underbrace{r(\boldsymbol{\beta}, \mathbf{x}_i(t))}_{\text{hazard ratio (relative risk)}}$$

We assume $r(\boldsymbol{\beta}, \mathbf{0}) = 1$, so the baseline hazard $\alpha_0(t)$ is the hazard for an individual with all covariates equal to zero

The common choice of relative risk function is

$$r(\boldsymbol{\beta}, \mathbf{x}_i(t)) = \exp\left(\boldsymbol{\beta}^T \mathbf{x}_i(t)\right) = \exp\left(\beta_1 x_{i1}(t) + \cdots + \beta_p x_{ip}(t)\right)$$

which gives Cox's regression model

e^{β_j} is the hazard ratio for one unit's increase in the *j*-th covariate, keeping the others constant (exercise 1.5)

Additive regression models

Hazard rate for individual i

$$\alpha(t \mid \mathbf{x}_i) = \underbrace{\beta_0(t)}_{\text{baseline hazard}} + \beta_1(t) x_{i1}(t) + \cdots + \underbrace{\beta_p(t)}_{\text{excess risk at time } t \text{ per unit's increase of } x_{ip}(t)} x_{ip}(t)$$

Note that the $\beta_j(t)$'s are regression functions

The additve regression model is a flexible nonparametric model that allows the effect of covatiates to change over time

However, the model does not constrain the hazard to be non-negative

A note on covariates

We assume that the intensity processes depend on the covariate processes

$$\mathbf{x}_i(t) = (x_{i1}(t), x_{i2}(t), \ldots, x_{ip}(t))^T \qquad i = 1, \ldots, n$$

Throughout we will assume that the covariate processes are predictable

This implies that:

- fixed covariates should be measured in advance (i.e. at time zero) and remain fixed throughout the study
- the values at time t of time-dependent covariates should be known "just before" time t

You should never let covariates depend on information from the future!

It is useful to distinguish between external (or exogenous) and internal (or endogenous) covariates

Examples of external covariates are:

- Fixed covariates
- Defined time-dependent covariates: the complete covariate path is given at the outset of the study (e.g. a person's age at study time t)
- Ancillary time-dependent covariates: the path of a stochastic process that is not influenced by the event being studied (e.g. observed level of air pollution)

Time-dependent covariates that are not external, are called internal

One example is biochemical markers measured for the individuals during follow-up

Interpretation of regression analyses with internal time-dependent covariates is not at all straightforward!