The Design of a Survival Study

The design of survival studies are usually based on the logrank test, and sometimes assumes the exponential distribution.

- As in standard designs, the power depends on
  - The Type I error (significance level)
  - The difference of interest, $\Delta$, under $H_A$
  - Sample size.

- A notable difference from the usual scenario is that power depends on the number of failures, not the total sample size.

- In practice, designing a survival study involves deciding how many patients or individuals to enter, as well as how long they should be followed.

- Designs may be for fixed sample size or sequential.
Review of power calculations for 2-sample normal

Suppose we have the following data:

Group 1: \((Y_{11}, \ldots Y_{1n_1})\)

Group 0: \((Y_{01}, \ldots Y_{0n_0})\)

and make the following assumptions:

\[ Y_{1j} \sim \mathcal{N}(\mu_1, \sigma^2) \quad Y_{0j} \sim \mathcal{N}(\mu_0, \sigma^2) \]

Our objective is to test:

\[ H_0 : \mu_1 = \mu_0 \quad \Rightarrow \quad H_0 : \Delta = 0 \quad \text{where} \quad \Delta = \mu_1 - \mu_0 \]

The standard test is based on the \(Z\) statistic:

\[ Z = \frac{\overline{Y}_1 - \overline{Y}_0}{\sqrt{\sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_0} \right)}} \]

where \(\sigma^2\) is often unknown in practice, and will eventually be estimated by \(s^2\), the pooled sample variance. For the purposes here we will assume the needed sample size will be large enough that the \(t\)-distribution can be approximated by \(\mathcal{N}(0, 1)\).

The test statistic \(Z\) follows a \(\mathcal{N}(0, 1)\) distribution under \(H_0\).
If the sample sizes are equal in the two arms, \( n_0 = n_1 = n/2 \), (which will maximize the power), then we have the simpler form:

\[
Z = \frac{\bar{Y}_1 - \bar{Y}_0}{\sqrt{s^2 \left( \frac{1}{n/2} + \frac{1}{n/2} \right)}} = \frac{\bar{Y}_1 - \bar{Y}_0}{2s/\sqrt{n}}
\]

The steps to follow in calculating the sample size are:

1. Determine the critical value, \( c \), for rejecting the null when it is true.
2. Calculate the probability of rejecting the null (i.e. power) when the alternative is true, substituting \( c \) from above.
3. Rewrite the expression in terms of the sample size for a given power.
Step (1):

Set the significance level, \( \alpha \), equal to the probability of rejecting the null hypothesis when it is true:

\[
\alpha = P \left( |\bar{Y}_1 - \bar{Y}_0| > c \mid H_0 \right)
\]

\[
= P \left( \frac{|\bar{Y}_1 - \bar{Y}_0|}{2\sigma/\sqrt{n}} > \frac{c}{2\sigma/\sqrt{n}} \mid H_0 \right)
\]

\[
= P \left( |Z| > \frac{c}{2\sigma/\sqrt{n}} \right) = 2 \cdot \Phi \left( -\frac{c}{2\sigma/\sqrt{n}} \right)
\]

so \( z_{1-\alpha/2} = \frac{c}{2\sigma/\sqrt{n}} \)

or \( c = \frac{z_{1-\alpha/2} \cdot 2 \cdot \sigma}{\sqrt{n}} \)

Note that \( z_\gamma \) is the value such that \( \Phi(z_\gamma) = Pr(Z < z_\gamma) = \gamma \).
Step (2):

Calculate the probability of rejecting the null when $H_A$ is true. Start out by writing down the probability of a Type II error:

$$
\beta = P(\text{accept } H_0 \mid H_a)
$$

so $1 - \beta = P(\text{reject } H_0 \mid H_a)$

$$
= P\left(\left|\overline{Y}_1 - \overline{Y}_0\right| > c \mid H_a\right)
$$

$$
= P\left(\frac{\left|\overline{Y}_1 - \overline{Y}_0\right| - \Delta}{2\sigma/\sqrt{n}} > \frac{c - \Delta}{2\sigma/\sqrt{n}} \mid H_a\right)
$$

$$
= P\left(Z > \frac{c - \Delta}{2\sigma/\sqrt{n}}\right)
$$

so we get

$$
z_\beta = -z_{1-\beta} = \frac{c - \Delta}{2\sigma/\sqrt{n}}
$$

Now we substitute $c$ from Step (1):

$$
-z_{1-\beta} = \frac{z_{1-\alpha/2}}{} \frac{2 \sigma/\sqrt{n} - \Delta}{2\sigma/\sqrt{n}}
$$

$$
= z_{1-\alpha/2} - \frac{\Delta}{2 \sigma/\sqrt{n}}
$$
Step (3):

Now rewrite the equation in terms of sample size for a given power, $1 - \beta$, and significance level, $\alpha$:

$$z_{1-\alpha/2} + z_{1-\beta} = \frac{\Delta}{2\sigma/\sqrt{n}}$$

$$= \frac{\Delta\sqrt{n}}{2\sigma}$$

$$\implies n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 4\sigma^2}{\Delta^2}$$

Notes:

The power is an increasing function of the ‘effect size’: $\Delta/\sigma$, which is the number of standard deviations between the two means, under the assumption of equal variances.

1. As $n$ increases, the power increases.
2. For fixed $n$ and $\sigma$, the power increases with $\Delta$.
3. For fixed $n$ and $\Delta$, the power decreases with $\sigma$.
4. Assigning equal numbers of patients to the two groups ($n_1 = n_0 = n/2$) is best in terms of maximizing power.
An Example:

\[ n = \left(\frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\Delta^2}\right)^2 \frac{4\sigma^2}{\Delta^2} \]

Say we want to derive the total sample size required to yield 90% power for detecting a difference of 0.5 standard deviations between means, based on a two-sided 0.05 level test.

\[ \alpha = 0.05 \]
\[ z_{1-\frac{\alpha}{2}} = 1.96 \]
\[ \beta = 0.10 \]
\[ z_{1-\beta} = z_{0.90} = 1.28 \]

\[ n = \frac{(1.96 + 1.28)^2 \sigma^2}{\Delta^2} \approx \frac{42 \sigma^2}{\Delta^2} \]

For a 0.5 standard deviation difference, \( \Delta/\sigma = 0.5 \), so

\[ n \approx \frac{42}{(0.5)^2} = 168 \]

If you end up with \( n < 30 \), then you should be using the \( t \)-distribution rather than the normal to calculate critical values, and then the process is iterative.
Survival Studies: Comparing Proportions of Events

In some cases, like in phase II cancer clinical trials where time to tumor progression is the primary endpoint, the sample size for a survival trial is based on a comparison of the proportion of events at some fixed point in time.

In this case, the event times are typically short and there is little/no censoring. Then the endpoint is really **binary** instead of ‘survival’ type, and we use normal approximations to the Binomial:

**Define:** (two-sample case)

- $P_c$ probability of event in control arm by time $t$
- $P_e$ probability of event in “experimental” arm by time $t$

The number of patients required per treatment arm based on a chi-square test comparing binomial proportions is:

$$N = \left\{ \frac{z_{1-\alpha/2} \sqrt{2\bar{P}(1-\bar{P})} + z_{1-\beta} \sqrt{P_e(1-P_e) + P_c(1-P_c)}}{(P_c - P_e)^2} \right\}^2$$

where $\bar{P} = (P_e + P_c)/2$
Notes on comparing proportions:

• Use of chi-square test is best when $0.2 < P_e, P_c < 0.8$

• Should have $\geq 15$ patients in each cell of the (2x2) table

• For smaller sample sizes, use Fisher’s exact test for power calculations.

• Efficiency compared to logrank test is near 100% for studies with short durations relative to the median event time;

• Calculation of sample size for comparing proportions often provides a rough idea about those based on comparison of survival distributions.
Sample size based on the logrank test

Recap: Consider a two group survival problem, with equal numbers of individuals in the two groups (say \( n_0 \) in group 0 and \( n_1 \) in group 1, \( n_0 = n_1 \)). Let \( t_1, \ldots, t_K \) represent the \( K \) ordered, distinct failure times, and at the \( j \)-th event time:

<table>
<thead>
<tr>
<th>Group</th>
<th>Die/Fail</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>( d_{0j} )</td>
<td>( r_{0j} - d_{0j} )</td>
<td>( r_{0j} )</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>( d_{1j} )</td>
<td>( r_{1j} - d_{1j} )</td>
<td>( r_{1j} )</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>( d_j )</td>
<td>( r_j - d_j )</td>
<td>( r_j )</td>
</tr>
</tbody>
</table>

where \( d_{0j} \) and \( d_{1j} \) are the number of events from group 0 and 1, respectively, at the \( j \)-th event time, and \( r_{0j} \) and \( r_{1j} \) are the corresponding numbers at risk.

The logrank test is: (z-statistic version)

\[
Z = \frac{\sum_{j=1}^{K}(d_{1j} - e_j)}{\sqrt{\sum_{j=1}^{K} v_j}}
\]

with \( e_j = d_j \cdot r_{1j}/r_j \)

\( v_j = r_{1j}r_{0j}d_j(r_j - d_j)/[r_j^2(r_j - 1)] \)
Distribution of the logrank statistic

Suppose that the hazard rates in the two groups are $\lambda_0(t)$ and $\lambda_1(t)$, with hazard ratio

$$\theta = e^\beta = \frac{\lambda_1(t)}{\lambda_0(t)}$$

We are interested in testing $H_0 : \beta = \log(\theta) = 0$

Assume that

- there are no ties;
- $\beta$, or equivalently $\theta$, is the true value that generates the data, i.e. under the alternative hypothesis (note that we assume the PH model);
- probability of assignment to the two groups are $P_0$, $P_1$ ($P_0 + P_1 = 1$).
Then conditional on the risk set, $d_{1j}$ is Bernoulli with mean (Schoenfeld, 1981)

\[
\begin{align*}
\frac{r_{1j}\lambda_1(t_j)}{r_{0j}\lambda_0(t_j) + r_{1j}\lambda_1(t_j)} &= \frac{p_{1j}\lambda_1(t_j)/\lambda_0(t_j)}{p_{0j} + p_{1j}\lambda_1(t_j)/\lambda_0(t_j)} \\
&= \frac{p_{1j}e^\beta}{p_{0j} + p_{1j}e^\beta} \\
&\approx \frac{p_{1j}(1 + \beta)}{p_{0j} + p_{1j}(1 + \beta)} \\
&= \frac{p_{1j} + p_{1j}\beta}{1 + p_{1j}\beta}
\end{align*}
\]

where $p_{0j} = r_{0j}/r_j$, $p_{1j} = r_{1j}/r_j$. Note the Taylor expansion assumes that $\beta$ is small.

Therefore

- $E(d_{1j} - e_j|d_{1j}, d_{0j}, r_{1j}, r_{0j}) \approx \frac{p_{1j} + p_{1j}\beta}{1 + p_{1j}\beta} - p_{1j} = p_{0j}p_{1j}\beta$
- also $v_j \approx p_{0j}p_{1j}$

Assuming that the two group have the same censoring distribution, also with $\beta$ small, eventually $p_{0j}$ and $p_{1j}$ are replaced by $P_0$ and $P_1$. So we have:

\[
E(Z) \approx \frac{\sum_{j=1}^{K} P_0 P_1 \beta}{\sqrt{\sum_{j=1}^{K} P_0 P_1}} = \sqrt{dP_0 P_1} \beta
\]

where $d$ is the total number of events, and $Z \sim N(\beta\sqrt{dP_0 P_1}, 1)$. 
Power of the Logrank Test

Using a similar derivation as for the reviewed normal distribution, the power of the logrank test for equal assignment to the two groups is approximately:

\[
\text{Power} \approx 1 - \Phi \left[ z_{1-\frac{\alpha}{2}} - \beta \sqrt{d/2} \right]
\]

**Note:** Power depends only on \(d\) and \(\beta\), not total sample size \(N\).

[Note: we will use \(\log(\theta)\) rather than \(\beta\) in the following, so that there is no confusion with the Type II error rate.]

To yield power = \(1 - \beta\), we want \(d\) so that

\[
1 - \beta = 1 - \Phi \left( z_{1-\frac{\alpha}{2}} - \log(\theta) \sqrt{d/2} \right)
\]

\[
\Rightarrow z_\beta = z_{1-\frac{\alpha}{2}} - \log(\theta) \sqrt{d/2}
\]

\[
\Rightarrow d = \frac{4 \left( z_{1-\frac{\alpha}{2}} - z_\beta \right)^2}{[\log(\theta)]^2}
\]

or

\[
\Rightarrow d = \frac{4 \left( z_{1-\frac{\alpha}{2}} + z_{1-\beta} \right)^2}{[\log(\theta)]^2}
\]

Note that the above argument can be easily extended to left truncated data (Xu and Chambers, 2011).
Example:

For a 2-arm study, to detect a hazard ratio of 1.5 with 90% power at a 2-sided significance level of $\alpha = 0.05$, we need the number of events:

$$d = \frac{4\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2}{[\log(\theta)]^2}$$

$$= \frac{4(1.96 + 1.282)^2}{[\log(1.5)]^2}$$

$$\approx \frac{42}{0.1644} = 256$$

# Events required for various Hazard Ratios

<table>
<thead>
<tr>
<th>Hazard Ratio</th>
<th>Power 80%</th>
<th>Power 90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>191</td>
<td>256</td>
</tr>
<tr>
<td>2.0</td>
<td>66</td>
<td>88</td>
</tr>
<tr>
<td>2.5</td>
<td>38</td>
<td>50</td>
</tr>
<tr>
<td>3.0</td>
<td>26</td>
<td>35</td>
</tr>
</tbody>
</table>

Many studies are designed to detect a hazard ratio of 1.5-2.0.
Practical Considerations

- How do we decide on $\theta$?
- How do we translate numbers of failures to numbers of patients?

We typically make the assumption of exponential distributions here.

**Median:**

If $T_i \sim \exp(\lambda_i)$, then

$$\text{Median}(T_i) = \log 2/\lambda_i$$

It follows that

$$\frac{\text{Median}(T_1)}{\text{Median}(T_0)} = \frac{\lambda_0}{\lambda_1} = \frac{1}{\theta}$$

Hence, doubling the median survival of a treated compared to a control group will correspond to halving the hazard.
**R-year survival rates**

Suppose the R-year survival rate in group 1 is $S_1(R)$ and in group 0 is $S_0(R)$. Under the exponential model:

$$S_i(R) = \exp(-\lambda_i R)$$

Hence,

$$\frac{\log(S_1(R))}{\log(S_0(R))} = \frac{-\lambda_1 R}{-\lambda_0 R} = \frac{\lambda_1}{\lambda_0} = \theta$$

Note that this result does not depend on R!.

**Example:** Suppose the 5-year survival rate on treatment A is 20% and we want 90% power to detect an improvement of that rate to 30%. The corresponding hazard ratio of treated to control is:

$$\frac{\log(0.3)}{\log(0.2)} = \frac{-1.204}{-1.609} = 0.748$$

From our previous formula, the number of events (deaths) needed to detect this improvement with 90% power, based on a 2-sided 5% level test is:

$$d = \frac{4(1.96 + 1.282)^2}{[\log(0.748)]^2} = 499$$
Translating to Number of Enrolled Patients

First, suppose that we will enter $N$ patients into our study at time 0, and will then continue the study for $F$ units of time.

Under $H_0$, the probability that an individual will fail during the study, again based on exponential assumption:

$$P(\text{fail}) = \int_0^F \lambda_0 e^{-\lambda_0 t} dt$$

$$= 1 - e^{-\lambda_0 F}$$

Hence

$$d = \left(\frac{N}{2}\right)(1 - e^{-\lambda_0 F}) + \left(\frac{N}{2}\right)(1 - e^{-\lambda_1 F})$$

To solve the above equation for $N$, we need to supply values of $F$ and $d$.

In general, the longer the follow-up, the fewer the patients that we need, and vice versa.
Example: Suppose we want to detect a 50% improvement in the median survival from 12 months to 18 months with 80% power at \( \alpha = 0.05 \), and we plan on following patients for 3 years (36 months).

\[
\text{Median}(T_i) = \log 2 / \lambda_i
\]

so \( \lambda_1 = \frac{\log 2}{M_1} = \frac{0.6931}{18} = 0.0385 \)

\( \lambda_0 = \frac{\log 2}{M_0} = \frac{0.6931}{12} = 0.0578 \)

\( \theta = \frac{\lambda_1}{\lambda_0} = \frac{0.0385}{0.0578} = \frac{12}{18} = 0.667 \)

and from our previous table, \# events required is \( d = 191 \) (same for \( \theta = 1.5 \) as it is for \( 1/1.5=0.667 \)).

So we need to solve:

\[
191 = \frac{N}{2}(1 - e^{-0.0578\times36}) + \frac{N}{2}(1 - e^{-0.0385\times36})
\]

\[
= \frac{N}{2}(0.875) + \frac{N}{2}(0.7500) = \frac{N}{2}(1.625)
\]

\( \Rightarrow \ N = 235 \)
A more realistic accrual pattern

In reality, not everyone will enter the study at time 0. Instead, the accrual will occur in a “staggered” manner over a period of time.

The standard assumption:
Suppose individuals enter the study uniformly over an accrual period lasting $A$ units of time, and that after the accrual period, follow-up will continue for another $F$ units of time.

To translate $d$ to $N$, we need to calculate the probability that a patient fails under this accrual and follow-up scenario.

$$P(\text{fail}) = \int_0^A P(\text{fail}|\text{enter at } A - a) \, f(a) \, da$$

$$= 1 - \frac{\int_0^A S(a + F) \, da}{A}$$

(1)

Then solve:
$$d = \left(\frac{N}{2}\right)P(\text{fail}; \lambda_0) + \left(\frac{N}{2}\right)P(\text{fail}; \lambda_1)$$
$$= \left(\frac{N}{2}\right)P_0 + \left(\frac{N}{2}\right)P_1$$
$$= \left(\frac{N}{2}\right)(P_0 + P_1)$$

$P_0$ and $P_1$ can be calculated once we assume some distribution, eg. $\text{Exp}(\lambda_i)$. We then solve for $N$. 

19
Other important issues:

The approaches just described address the basic question of calculating a sample size for study with a survival endpoint.

These approaches often need to be modified to address the following complications:

- Loss to follow-up
- Non-compliance (or cross-overs)
- Stratification
- Sequential monitoring
- Equivalence hypotheses
Loss to follow-up

If some patients are lost to follow up (as opposed to censored at the end of the trial without the event), the power will be decreased.

- **Simple inflation method** - If $\ell\times100\%$ of patients are anticipated to be lost to follow up, calculate target sample size to be

$$N^* = \left(\frac{1}{1 - \ell}\right) \cdot N$$

**Example:** Say you calculate $N = 200$, and anticipate losses of 20%. The simple inflation method would give you a target sample size of $N^* = (1/0.8)\times200 = 250$.

**Warning:** people often make the mistake of just inflating the original sample size by $\ell\times100\%$, which would have given $N^* = 240$ for the example above.

- **Exponential loss assumption** - the above approach assumes that losses contribute no information. But we actually have information on them up until the time that they are lost. People sometimes incorporate this by assuming that time to loss also follows an exponential distribution, and modify $P_0$ and $P_1$. 
Sequential Design

In clinical trials and other studies, it is often desirable to conduct interim analyses of a study while it is still ongoing.

Rationale:

- **ethical**: if one treatment is substantially worse than another, then it is wrong to continue to give the inferior treatment to patients.

- **timely reporting**: if the hypothesis of interest has been clearly established halfway through the study, then science and the public may benefit from early reporting.

**WARNING:**
Unplanned interim analyses (i.e. multiple testing) can seriously inflate the true type I error of a trial. If interim analyses are to be performed, it is important to carefully plan these in advance, and to adjust all tests appropriately so the type I error is of the desired size.
The table below shows the Type I error rate if each test is done at $\alpha = 0.05$ for various numbers of interim analyses:

<table>
<thead>
<tr>
<th>Number of interim analyses</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>10</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>8.3%</td>
<td>10.7%</td>
<td>12.6%</td>
<td>14.2%</td>
<td>19.3%</td>
<td>26.6%</td>
<td></td>
</tr>
</tbody>
</table>

(from Lee, *Statistical Methods for Survival Data*, Table 12.9)

What can we do to protect against this type I error inflation?

**Pocock Approach:**

Pick a smaller significance level (say $\alpha'$) to use at each interim analysis so that the overall type I error stays at level $\alpha$.

A problem with the Pocock method is that even the very last analysis has to be performed at level $\alpha'$. This tends to be very conservative at the final analysis.

**O’Brien and Fleming Approach:**

A preferable approach would be to vary the alpha levels used for each of the K interim analyses, and try to keep the very last one “close” to the desired overall significance level. The O’Brien-Fleming approach does that.
Comments and notes:

• There are several other approaches available for sequential design and analysis. The O’Brien and Fleming approach is probably the most popular in practice.

• There are many variations on the theme of sequential design. The type we have discussed here is called **group sequential analysis**.
  – There are other approaches that require continuous monitoring/analysis after each new individual enters the study!
  – An example of this is in phase I clinical trials using the continual reassessment methods (CRM).

• Some designs allow for early stopping in the absence of a sufficient treatment effect or, ‘accepting the null’. These procedures often use “stochastic curtailment” or “conditional power” calculations.

• In large randomized Phase III studies we often have 1-3 interim looks, at equal ‘information’ times (i.e. roughly equal number of events between two consecutive looks).
• Going from a fixed to a group sequential design usually adds only a small portion to the required sample size. This would give you some idea in calculating the sample size when you plan on doing interim monitoring.

• “Non-statistical” issues such as safety may enter decisions about whether or not to stop a trial early.

A useful reference book is Jennison and Turnbull ‘Group Sequential Methods’.