Chapter 11

Designing a Survival Study

We will focus on the power of tests based on the exponential distribution and the logrank test.

- As in standard designs, the power depends on
  - The Type I error (significance level)
  - The difference of interest, Δ, under H_a.
- 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 **fixed sample size** or **sequential**
  (More on this later!)

References:

Collett	Chapter 12
Pocock	Chapter 9 of *Clinical Trials*
Williams	Chapter 10 of *AIDS Clinical Trials*
	(eds. Finkelstein and Schoenfeld)

11.1 Review of power calculations for 2-sample normal

Suppose we have the following data:

$$\begin{array}{ll} \text{Group 1:} & (Y_{11}, \ldots Y_{1n_1}) \\ \text{Group 0:} & (Y_{01}, \ldots Y_{0n_0}) \end{array}$$

and make the following assumptions:

$$Y_{1j} \sim \mathcal{N}(\mu_1, \sigma^2) \qquad Y_{0j} \sim \mathcal{N}(\mu_0, \sigma^2)$$

Our objective is to test:

$$H_0 : \mu_1 = \mu_0 \quad \Rightarrow \quad H_0 : \triangle = 0 \qquad \text{where } \triangle = \mu_1 - \mu_0$$

The standard test is based on the Z statistic:

$$Z = \frac{\overline{Y_1} - \overline{Y_0}}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_0})}}$$

where s^2 is the pooled sample variance (we are assuming equal variances here). This test statistic 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:

$$Z \;=\; \frac{\overline{Y_1} - \overline{Y_0}}{\sqrt{s^2(\frac{1}{n/2} + \frac{1}{n/2})}} = \frac{\overline{Y_1} - \overline{Y_0}}{2s/\sqrt{n}}$$

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 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, α, equal to the probability of rejecting the null hypothesis when it is true:

$$\begin{aligned}
\alpha &= Pr\left(|\overline{Y_1} - \overline{Y_0}| > c \mid H_0\right) \\
&= Pr\left(\frac{|\overline{Y_1} - \overline{Y_0}|}{2s/\sqrt{n}} > \frac{c}{2s/\sqrt{n}} \mid H_0\right) \\
&= Pr\left(|Z| > \frac{c}{2s/\sqrt{n}}\right) \;=\; 2 \cdot \Phi\left(\frac{c}{2s/\sqrt{n}}\right)
\end{aligned}$$

$$\begin{aligned}
\text{so} \quad z_{1-\alpha/2} &= \frac{c}{2s/\sqrt{n}} \\
\text{or} \quad c &= \frac{z_{1-\alpha/2}\ 2\ s}{\sqrt{n}}
\end{aligned}$$

Note that z_γ 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:

$$\begin{aligned}
\beta &= Pr\left(\text{accept } H_0 \mid H_a\right) \\
\text{so} \quad 1-\beta &= Pr\left(\text{reject } H_0 \mid H_a\right) \\
&= Pr\left(|\overline{Y_1} - \overline{Y_0}| > c \mid H_a\right) \\
&= Pr\left(\frac{|\overline{Y_1} - \overline{Y_0}| - \Delta}{2s/\sqrt{n}} > \frac{c-\Delta}{2s/\sqrt{n}} \mid H_a\right) \\
&= Pr\left(Z > \frac{c-\Delta}{2s/\sqrt{n}}\right)
\end{aligned}$$

$$\text{so we get} \quad z_\beta = -z_{1-\beta} = \frac{c-\Delta}{2s/\sqrt{n}}$$

Now we substitute c from Step (1):

$$\begin{aligned} -z_{1-\beta} &= \frac{z_{1-\alpha/2}\ 2\ s/\sqrt{n} - \Delta}{2s/\sqrt{n}} \\ &= z_{1-\alpha/2} - \frac{\Delta}{2\ s\ /\sqrt{n}} \end{aligned}$$

Step (3):

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

$$\begin{aligned} z_{1-\alpha/2} + z_{1-\beta} &= \frac{\Delta}{2s/\sqrt{n}} \\ &= \frac{\Delta\sqrt{n}}{2s} \\ \Longrightarrow \quad n &= \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 4s^2}{\Delta^2} \end{aligned}$$

11.1.1 Notes

The power is an increasing function of the standardized difference:

$$\mu_T(\triangle) = \frac{\triangle}{2s/\sqrt{n}}$$

This is just the number of standard errors between the two means, under the assumption of equal variances.

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

11.1.2 An Example

$$n = \frac{\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2 4s^2}{\triangle^2}$$

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.

$$\begin{aligned} \alpha &= 0.05 \\ z_{1-\frac{\alpha}{2}} &= 1.96 \\ \beta &= 0.10 \\ z_{1-\beta} = z_{0.90} &= 1.28 \end{aligned}$$

$$n = \frac{(1.96 + 1.28)^2 \, 4s^2}{\triangle^2} \approx \frac{42 \; s^2}{\triangle^2}$$

For a 0.5 standard deviation difference, $\Delta/s = 0.5$, 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.

11.2 Survival Studies: Comparing Proportions of Events

In some cases, the sample size for a survival trial is based on a crude comparison of the proportion of events at some fixed point in time.

In this case, we can apply the results just shown to get sample sizes, based on the normal approximation to the binomial:

Define:

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:

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

where $\overline{P} = (P_e + P_c)/2$

This looks slightly different because the variance is not the same.

11.2.1 Notes on comparing proportions of failures

- Use of chi-square test is best when $0.2 < P_e, P_c < 0.8$
- Should have ≥ 15 patients in each cell of the (2x2) table
- For smaller sample sizes, use Fisher's exact test to motivate power calculations
- Efficiency vs logrank test is near 100% for studies with short durations relative to the median event time **What does this mean in terms of the event rates? High or low?**
- Calculation of sample size for comparing proportions often provides an upper bound to those based on comparison of survival distributions

11.3 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). Let $\tau_1, ..., \tau_K$ represent the K ordered, distinct failure times, and at the j-th event time:

	Die/Fail		
Group	Yes	No	Total
0	d_{0j}	$r_{0j} - d_{0j}$	r_{0j}
1	d_{1j}	$r_{1j} - d_{1j}$	r_{1j}
Total	d_j	$r_j - d_j$	r_j

where d_{0j} and d_{1j} are the number of deaths (events) in 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)

$$Z_{LR} = \frac{\sum_{j=1}^{K}(d_{1j} - e_j)}{\sqrt{\sum_{j=1}^{K} v_j}}$$

$$\text{with} \quad e_j = d_j \, r_{1j}/r_j$$

$$v_j = r_{1j} r_{0j} d_j (r_j - d_j)/[r_j^2 (r_j - 1)]$$

11.3.1 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

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

and suppose we are interested in testing $H_o : \beta = ln(\theta) = 0$
(which is equivalent to testing $H_o : \theta = 1$.)

[Note: I will use $ln(\theta)$ rather than β in the following, so that there is no confusion with the Type II error rate]

It is possible to show that

- if there are no ties, and
- we are "near" H_0 :

then:

- $E(d_{1j} - e_j | d_{1j}, d_{0j}, r_{1j}, r_{0j}) \approx ln(\theta)/4$
- $v_j \approx 1/4$

So, at a point $ln(\theta)$ in the alternative, we get:

$$Z_{LR} \approx \frac{\sum_{j=1}^{K} ln(\theta)/4}{\sqrt{\sum_{j=1}^{K} 1/4}} = \frac{d\ ln(\theta)/4}{\sqrt{d/4}} = \frac{\sqrt{d}\ ln(\theta)}{2}$$

$$\text{and } Z_{LR} \sim N(\ ln(\theta)\sqrt{d}/2\ ,\ 1\ )$$

Heuristic Proof:

$$\begin{aligned} E(d_{1j} | d_{1j}, d_{0j}, r_{1j}, r_{0j}) &= Pr(d_{1j} = 1 | d_j = 1, r_{1j}, r_{0j}) \\ &= \frac{r_{1j}\lambda_0\theta}{r_{1j}\lambda_0\theta + r_{0j}\lambda_0} \\ &= \frac{r_{1j}\theta}{r_{1j}\theta + r_{0j}} \\ &= \frac{r_{1j}}{r_{1j} + r_{0j}} + ln(\theta)\left[\frac{r_{1j}r_{0j}}{(r_{1j} + r_{0j})^2}\right] \end{aligned}$$

But $e_j = r_{1j}/(r_{1j} + r_{0j})$, so:

$$E(d_{1j}|d_{1j}, d_{0j}, r_{1j}, r_{0j}) - e_j = ln(\theta) \left[\frac{r_{1j} r_{0j}}{(r_{1j} + r_{0j})^2} \right]$$

If $n_0 = n_1$, then near H_0 :, $r_{1j} \approx r_{0j}$, hence,

$$E(d_{1j}|d_{1j}, d_{0j}, r_{1j}, r_{0j}) - e_j = ln(\theta)/4$$

Similarly, with no ties, we have

$$v_j = r_{1j} r_{0j} / r_j^2 \approx 1/4$$

This can also be derived via the partial likelihood:
(if you're interested)

We can write the partial likelihood as:

$$\begin{aligned} l(\beta) &= \log \left[\prod_{j=1}^{n} \left(\frac{e^{\beta \mathbf{Z}_j}}{\sum_{\ell \in \mathcal{R}(\tau_j)} e^{\beta \mathbf{Z}_\ell}} \right)^{\delta_j} \right] \\ &= \sum_{j=1}^{n} \delta_j \left[\beta \mathbf{Z}_j - \log \left(\sum_{\ell \in \mathcal{R}(\tau_j)} e^{\beta \mathbf{Z}_\ell} \right) \right] \end{aligned}$$

and then the **"score"** (partial derivative of log-likelihood) becomes:

$$\begin{aligned} U(\beta) &= \frac{\partial}{\partial \beta} \ell(\beta) \\ &= \sum_{j=1}^{n} \delta_j \left[\mathbf{Z}_j - \frac{\sum_{\ell \in \mathcal{R}(\tau_j)} \mathbf{Z}_\ell \; e^{\beta \mathbf{Z}_\ell}}{\sum_{\ell \in \mathcal{R}(\tau_j)} e^{\beta \mathbf{Z}_\ell}} \right] \end{aligned}$$

We can write the **"information"** (minus second partial derivative of the log-likelihood) as:

$$-\frac{\partial^2}{\partial \beta^2} \ell(\beta) = \sum_{j=1}^{n} \delta_j \left[\frac{\sum_{\ell \in \mathcal{R}(\tau_j)} e^{\beta \mathbf{Z}_\ell} \sum_{\ell \in \mathcal{R}(\tau_j)} \mathbf{Z}_\ell e^{\beta \mathbf{Z}_\ell} - (\sum_{\ell \in \mathcal{R}(\tau_j)} \mathbf{Z}_\ell e^{\beta \mathbf{Z}_\ell})^2}{\sum_{\ell \in \mathcal{R}(\tau_j)} e^{\beta \mathbf{Z}_\ell}} \right]$$

The logrank statistic (with no ties) is equivalent to the score statistic for testing $\beta = 0$:

$$Z_{LR} \quad \frac{U(0)}{\sqrt{I(0)}}$$

By a Taylor series expansion:

$$U(0) \cong U(\beta) - \beta \frac{\partial U}{\partial \beta}(0)$$

$$E[U(0)] \cong \beta d/4 \qquad \text{and} \quad I(0) \cong d/4$$

11.4 Power of the Logrank Test

Using a similar argument to before, the power of the logrank test (based on a two-sided α level test) is approximately:

$$\text{Power}(\theta) \approx 1 - \Phi\left[z_{1-\frac{\alpha}{2}} - ln(\theta)\sqrt{d}/2\right]$$

Note: Power depends only on d and θ!

We can easily solve for the required number of events to achieve a certain power at a specified value of θ:

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

$$\begin{aligned} 1-\beta &= 1 - \Phi\left(z_{1-\frac{\alpha}{2}} - ln(\theta)\sqrt{d}/2\right) \\ \Rightarrow \quad z_\beta &= z_{1-\frac{\alpha}{2}} - ln(\theta)\sqrt{d}/2 \\ \Rightarrow \quad d &= \frac{4\left(z_{1-\frac{\alpha}{2}} - z_\beta\right)^2}{[ln(\theta)]^2} \\ \text{or} \quad d &= \frac{4\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2}{[ln(\theta)]^2} \end{aligned}$$

11.4.1 Example:

Say we were planning a 2-arm study, and wanted to be able to detect a hazard ratio of 1.5 with 90% power at a 2-sided significance level of $\alpha = 0.05$.

Required number of events:

$$\begin{aligned} d &= \frac{4\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2}{[ln(\theta)]^2} \\ &= \frac{4(1.96 + 1.282)^2}{[ln(1.5)]^2} \\ &\approx \frac{42}{0.1644} = 256 \end{aligned}$$

Most studies are designed to detect a hazard ratio of 1.5-2.0.

Table 11.1: Number of events required for various Hazard Ratios

Hazard	**Power**	
Ratio	80%	90%
1.5	191	256
2.0	66	88
2.5	38	50
3.0	26	35

11.4.2 Practical Considerations

- How do we decide on θ?
- How do we translate numbers of failures to numbers of patients?

11.4.3 Hazard ratios for the exponential distribution

The hazard ratio from two exponential distributions can be easily translated into more intuitively interpretable quantities:

Median:

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

$$\text{Median}(T_i) = -ln(0.5)/\lambda_i$$

It follows that

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

Hence, doubling the median survival of a treated compared to a control group will correspond to halving the hazard.

11.4.4 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,

$$\frac{ln(S_1(R))}{ln(S_0(R))} = \frac{-\lambda_1 R}{-\lambda_0 R} = \frac{\lambda_1}{\lambda_0} = e^{\beta} = \theta$$

Hence, doubling the hazard rate from group 1 to group 0 will correspond to doubling the log of the R-year survival rate. Note that this result does not depend on R!.

11.4.5 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:

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

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:

$$d = \frac{4(1.96 + 1.282)^2}{[ln(0.748)]^2} = 499$$

11.4.6 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 is:

$$Pr(fail) = \int_0^F \lambda_0 e^{-\lambda_0 t} dt = 1 - e^{-\lambda_0 F}$$

If we need d failures, then to calculate the sample size we solve

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

To solve the above equation for N, we need to supply values of F and d. In other words, here we are already deciding what HR we want to detect (with what power, etc), and for how long we are going to follow patients. What we get is the total number of patients we need to enroll in order to observe the desired number of events in F units of follow-up time.

11.4.7 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).

We can use the two medians to calculate both the parameters λ_0 and λ_1 and the hazard ratio, θ:

$$\begin{aligned} \text{Median}(T_i) &= -ln(0.5)/\lambda_i \\ \text{so } \lambda_1 &= \frac{-ln(0.5)}{M1} = \frac{0.6931}{18} = 0.0385 \\ \lambda_0 &= \frac{-ln(0.5)}{M0} = \frac{0.6931}{12} = 0.0578 \\ \theta &= \frac{\lambda_1}{\lambda_0} = \frac{0.0385}{0.0578} = \frac{12}{18} = 0.667 \end{aligned}$$

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:

$$\begin{aligned} 191 &= (N/2)(1 - e^{-0.0578*36}) + (N/2)(1 - e^{-0.0385*36}) \\ &= (N/2)(0.875) + (N/2)(0.7500) = (N/2)(1.625) \\ \Rightarrow \quad N &= 235 \end{aligned}$$

(for practical reasons, we would probably round up to 236 and randomize 118 patients to each treatment arm)

11.4.8 More realistic accrual patterns

In reality, not everyone will enter the study on the same day. 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.

$$\begin{aligned} Pr(\text{fail}) &= \int_0^A Pr(\text{fail}|\text{enter at } a)\ f(a)\ da \\ &= 1 - \frac{\int_0^A S(a+F)\ da}{A} \end{aligned} \tag{11.1}$$

$$\begin{aligned} \text{Then solve:} \quad d &= (N/2)Pr(\text{fail};\lambda_0) + (N/2)Pr(\text{fail};\lambda_1) \\ &= (N/2)P_c + (N/2)P_e \\ &= (N/2)(P_c + P_e) \end{aligned}$$

If we now solve for N (substituting in formula for d), we get:

$$N = \frac{2\ d}{(P_c + P_e)}$$

$$N = \frac{8\ \left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2}{[ln(\theta)]^2}\ \ \frac{1}{(P_c + P_e)}$$

How can we get P_c and P_e from (11.1)?

If we assume that the exponential distribution holds, then we can solve (11.1) to obtain:

$$P_i = 1 - \frac{exp(-\lambda_i F)(1 - exp(-\lambda_i A))}{\lambda_i A} \tag{11.2}$$

(for $i = c, e$)

Freedman suggested an approximation for P_c and P_e, by computing the probability of an event at the median duration of follow-up, $(A/2 + F)$:

$$P_i = Pr(\text{fail};\lambda_i) = 1 - exp[-\lambda_i(A/2 + F)] \tag{11.3}$$

He showed that this approximation works pretty well for the exponential distribution (i.e., it gives values close to (11.2)).

11.4.9 An alternative formulation

Rubenstein, Gail, and Santner (1981) suggest the following approach for calculating the total sample size that must be enrolled:

$$N = \frac{2\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2}{[ln(\theta)]^2}\left[\frac{1}{P_c} + \frac{1}{P_e}\right]$$

where P_c and P_e are the expected proportion of patients or individuals who will fail (have an event) on the control and treatment arms.

How do we calculate (estimate) P_c and P_e ?

- using the general formula for a distribution S given in (11.1)
- using the exact formula for an exponential distribution given in (11.2)
- using the approximation given by (11.3)

Note: all of these formulas can be modified for **unequal** assignment to treatment (or exposure) groups by changing $(N/2)$ in the formulas on p.17-19 to $(qc * N)$ and $(qe * N)$, where qc and qe are the proportions assigned to the control and exposed groups, respectively.

11.4.10 Freedman's Approach (1982)

Freedman's approach is based on the logrank statistic under the assumption of proportional hazards, but does <u>not</u> require the assumption of exponential survival distributions.

Total number of events:

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

Total sample size:

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

where P_e and P_c are estimated using (11.3).

This approximation depends on the assumption of a constant ratio between the number of patients at risk in the two treatment groups prior to each event time $\Longrightarrow r_{0j} \approx r_{1j}$ (as

shown in the "heuristic proof"). When this assumption is not satisfied, the required sample sizes tend to be overestimated.

Q. When would this assumption not be satisfied?

A. When the smallest detectable difference is large

11.4.11 Some examples of study design

Example I:

A clinical trial in esophageal cancer will randomize patients to radiotherapy alone (Rx A) versus radiotherapy plus chemotherapy (Rx B). The goal of the study is to compare the two treatments with respect to survival, and we plan to use the logrank test. From historical data, we know that the median survival on RX A for this disease is around 9 months. We want 90% power to detect an improvement in this median to 18 months. Past studies have been able to accrue approximately 50 patients per year. Choose a suitable study design.

Example II:

A clinical trial in early stage breast cancer will randomize patients after their surgery to Tamoxifen (ARM A) versus observation only (ARM B). The goal of the study is to compare the two treatments with respect to time to relapse, and the logrank test will be used in the analysis. From historical data, we know that after five years, 65% of the patients will still be disease free. We would like to have 90% power to detect an improvement in this disease free rate to 75%. Past studies have been able to accrue approximately 200 patients per year. Choose a suitable study design.

Example III:

Some investigators in the environmental health department want to conduct a study to assess the effects of exposure to toluene on time to pregnancy. They will conduct a cohort study involving women who work in a chemical factory in China. It is estimated that 20% of the women will have workplace exposure to toluene. Furthermore, it is known that among unexposed women, 80% will become pregnant within a year. The investigators will be able to enroll 200 women per year into the study, and plan an additional year of follow-up at the end of accrual. Assuming they have 2 years accrual, what reduction in the 1-year pregnancy rate for exposed women will they be able to detect with 85% power? What if they have 3 years of accrual?

11.5 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 slightly to address the following complications:

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

For more information and references on these issues, see my book chapter from *AIDS Clinical Trials*. I will briefly summarize some of the main points.

11.5.1 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.

There are two main approaches for dealing with this:

- **Simple inflation method** - If ℓ*100% 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 $N^* = (1/0.8) * 200 = 250$.

  Warning: people often make the mistake of just inflating the original sample size by ℓ*100%, 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. Incorporate this by assuming that time to loss also follows an exponential distribution, and modify P_e and P_c (as described in my book chapter, p.151).

11.5.2 Noncompliance

If some patients don't take their assigned treatments, the power will be decreased. This issue has two sides:

- **Drop-outs (d_e)** - patients who cannot tolerate the medication stop taking it; their hazard rate would become the same as the placebo group (if included in study) at that point.
- **Drop-ins (d_c)** - patients assigned to less effective therapy may not get relief from symptoms and seek other therapy, or request to cross-over.

A conservative remedy – adjust P_e and P_c as follows:

$$\begin{aligned} P_e^* &= P_e(1-d_e) + P_c\ d_e \\ P_c^* &= P_c(1-d_c) + P_e\ d_c \end{aligned}$$

Design Strategy:

1. Decide on
   - Type I error (significance level)
   - clinically important difference (in terms of HR)
   - desired power
2. Determine the number of failures needed
3. Based on past experience
   - decide on a reasonable distribution for the controls (usually exponential)
   - estimate anticipated accrual per unit time
   - estimate expected rate of loss to follow up

   Vary the values of A and F until you get something practically feasible that gives the right number of failures.
4. Consider noncompliance, seq. monitoring, and other issues

Included on the next several pages is a SAS program I wrote to calculate sample sizes for survival studies. It uses several of the approaches we've discussed, including:

- Rubenstein, Gail and Santner (RGS, 1981)

- Freedman (1982)
- Lachin and Foulkes (1986)

A copy of this program is shown on the next several pages. The program requires entry of:

- Significance level (alpha)
- Power
- Sides (1 for one-sided test, 2 for two-sided test)
- Accrual period
- Follow up period
- Yearly rate of loss to follow-up
- Proportion randomized to experimental treatment arm
- One of the following:
  - Yearly event rate on control and exp. treatment arms
  - Yearly event rate on control arm, and the hazard ratio
  - Median time to event on control and exp. treatment arms

11.6 The SAS program rgsnew.sas

```
data rgs;
  ************************************************************************;
  *** enter the following information in this block;
  alpha = 0.05;          /* significance level */
  sides = 2;             /* one-sided or two-sided test */
  power = 0.90;          /* Desired power */
  accrual = 2;           /* Accrual period in years */
  fu = 1.5;              /* Follow up after last patient is accrued */
  loss = 0.0;            /* yearly rate of loss */
  qe = 0.5;              /* proportion randomized to experimental arm */

  *** either enter the median time to event in years on control;
  *** or the yearly event rate - leave the other value missing;
  medianc = 0.75;        /* median time to event on control arm */
  probc = .;             /* yearly event rate in control arm */

  *** either enter the yearly event rate in the experimental arm  ;
  *** or the hazard ratio for control vs experimental             ;
  *** or the median time to event on experimental arm in years    ;
  *** leave the other values missing (.);
  mediane = 1.5;         /* median time to event on experimental */
  probe = .;             /* yearly event rate in experimental arm */
  rr=.;                  /* hazard ratio */
  ************************************************************************;
  beta = 1 - power;
  qc = 1 - qe;
  zalpha = probit(1-alpha/sides);
  zbeta = probit(1-beta);

  *** calculate yearly event rate in both arms using medians, if supplied;
  if medianc^=. then do;
    hazc=-log(0.5)/medianc;
    probc=1-exp(-hazc);
  end;
  if mediane^=. then do;
    haze=-log(0.5)/mediane;
    probe=1-exp(-haze);
  end;

  hazc = -log(1-probc);

  *** calculate hazard in experimental group, using yearly event rate;
  *** or hazard ratio;
```

```
if probe^=. then haze = -log(1-probe);
if rr^=. then haze = hazc/rr;

if probe^=. and haze^=. then do;
   put "************************************************************";
   put "WARNING: both yearly event rate and hazard ratio (HR) have";
   put "         been specified.  Calculations will use the HR";
   put "************************************************************" /;
end;

*** calculate median survival times if not supplied;
medianc=-log(0.5)/hazc;
mediane=-log(0.5)/haze;

hazl = -log(1-loss);
avghaz = qc*hazc + qe*haze;
rr = hazc/haze;
log_rr = log(hazc/haze);
totloss=(accrual*0.5 + fu)*loss;

*** compute expected probability of death (event) during trial;
*** given staggered accrual but NO loss;
pc0loss = 1-((exp(-hazc*fu)-exp(-hazc*(accrual+fu)))/(hazc*accrual));
pe0loss = 1-((exp(-haze*fu)-exp(-haze*(accrual+fu)))/(haze*accrual));

*** compute expected probability of event during trial;
*** given staggered accrual AND loss;
pc = (1 - (exp(-(hazc+hazl)*fu)-exp(-(hazc+hazl)*(fu+accrual)))
          /((hazc+hazl)*accrual))*(hazc/(hazc+hazl));
pe = (1 - (exp(-(haze+hazl)*fu)-exp(-(haze+hazl)*(fu+accrual)))
          /((haze+hazl)*accrual))*(haze/(haze+hazl));
pbar = (1 - (exp(-(avghaz+hazl)*fu)-exp(-(avghaz+hazl)*(fu+accrual)))
          /((avghaz+hazl)*accrual))*(avghaz/(avghaz+hazl));

*** compute total sample size assuming loss;
N = int(((zalpha+zbeta)**2)/(log_rr**2)*(1/(qc*pc)+1/(qe*pe))) + 1;

*** Compute sample size using method of Freedman (1982);
N_FRD = int((2*(((rr+1)/(rr-1))**2)*(zalpha+zbeta)**2)/(2*(qe*pe+qc*pc)))+1;


*** Compute sample size using method of Lachin and Foulkes (1986);
*** with rates under H0 given by pooled hazard;
N_LF = int((zalpha*sqrt((avghaz**2)*(1/pbar)*(1/qc + 1/qe)) +
       zbeta*sqrt((hazc**2)*(1/(qc*pc)) + (haze**2)*(1/(qe*pe))))**2/
```

```
        ((hazc-haze)**2)) + 1;

  *** compute total sample size assuming no loss;
  n_0loss = int(((zalpha+zbeta)**2)/(log_rr**2)*
            (1/(qc*pc0loss)+1/(qe*pe0loss))) + 1;

  *** compute sample size using simple inflation method for loss;
  naive = int(n_0loss/(1-totloss)) + 1;

  *** verify that actual power is same as desired power;
  newpower = probnorm(sqrt(((N)*(log_rr**2))/(1/(qc*pc) + 1/(qe*pe)))
             - zalpha);

  if abs(newpower-power)>0.001 then do;
    put '*** WARNING: actual power is not equal to desired power';
    put 'Desired power: ' power ' Actual power: ' newpower;
  end;

  ************************************************************************;
  *** compute number of events expected during trial;
  ************************************************************************;
  *** Compute expected number under null;
  n_evth0 = int(n*pbar) + 1;

  r=qc/qe;
  *** Rubinstein, Gail and Santner (1981) method - simple approximation;
  n_evtrgs = int((((r+1)**2)/r)*((zalpha + zbeta)**2)/(log_rr**2)) + 1;

  *** Freedman (1982);
  n_evtfrd = int((((rr+1)/(rr-1))**2) * (zalpha+zbeta)**2)+1;

  *** Using backtracking method of Lachin and Foulkes (1986);
  n_evt_c = int(N*qc*pc) + 1;
  n_evt_e = int(N*qe*pe) + 1;
  n_evtlf = n_evt_c + n_evt_e;


label sides='Sides'
      alpha='Alpha'
      power='Power'
      beta='Beta'
      zalpha='Z(alpha)'
      zbeta='Z(beta)'
      accrual='Accrual (yrs)'
      fu='Follow-up (yrs)'
```

```
      loss='Yearly Loss'
      totloss='Total Loss'
      probc='Yearly event rate: control'
      probe='Yearly event rate: active'
      rr='Hazard ratio'
      log_rr='Log(HR)'
      N='Total Sample size (RGS)'
      N_FRD='Total Sample size (Freedman)'
      N_LF='Total Sample size (L&F)'
      n_0loss='Sample size (no loss)'
      pc='Pr(event), control'
      pe='Pr(event), active'
      pc0loss='Pr(event| no loss), control'
      pe0loss='Pr(event| no loss), active'
      n_evth0='# events (Ho-pooled)'
      n_evtrgs='# events (RGS)'
      n_evtfrd='# events (Freedman)'
      n_evtlf='# events (L&F)'
      medianc='Median survival, control'
      mediane='Median survival, active'
      naive='Sample size (naive loss)';


proc print data=rgs label noobs;
title 'Sample size & expected events for comparing two survival distributions';
title2 'Using method of Rubinstein, Gail and Santer (RGS, 1981)';
title3 'Freedman (1982), or Lachin and Foulkes (L&F, 1986)';
  var sides alpha power accrual fu loss totloss
        probc probe medianc mediane pc pe pc0loss pe0loss rr log_rr
        n_evth0 n_evtrgs n_evtfrd n_evtlf N N_FRD N_LF n_0loss naive;
  format power loss totloss f4.2 medianc mediane f5.3
         probc probe rr log_rr pc pe pc0loss pe0loss f6.4;
```

Back to Example I:

A clinical trial in esophageal cancer will randomize patients to radiotherapy alone (Rx A) versus radiotherapy plus chemotherapy (Rx B). The goal of the study is to compare the two treatments with respect to survival, and we plan to use the logrank test. From historical data, we know that the median survival on Rx A for this disease is around 9 months. We want 90% power to detect an improvement in this median to 18 months. Past studies have been able to accrue approximately 50 patients per year. Choose a suitable study design.

First, let's write down what we know:

- desired significance level not stated, so use $\alpha = 0.05$
  (assume a two-sided test)
- assume equal randomization to treatment arms
  (unless otherwise stated)
- desired power is 90%
- median survival on control is 9 months $\Rightarrow M0 = 9$
- want to detect improvement to 18 months on Rx B $\Rightarrow M1 = 18$
- Maximum accrual per year is 50 patients

We have all of the information we need to run my program, except the accrual and follow up times. We need to use trial and error to get these.

Accrual Period	Follow-up Period	Number of Events Required	Total Sample Size	Total Study Duration
1	2.5	88	106	3.5
2	1.5	88	115	3.5
2.5	1	88	122	3.5
3	0.5	88	133	3.5
3	1	88	117	4

Shown on the next page is the output I got from `rgsnew.sas` using Accrual=2, Follow-up=1.5. I've given the RGS numbers above.

Which of the above are feasible designs?

```
  Sample size & expected events for comparing two survival distributions
         Using method of Rubinstein, Gail and Santer (RGS, 1981)
            Freedman (1982), or Lachin and Foulkes (L&F, 1986)
                                                                  Yearly
                                                                   event
                         Accrual   Follow-up   Yearly   Total    rate:
Sides   Alpha   Power     (yrs)      (yrs)      Loss     Loss    control

  2     0.05    0.90        2         1.5       0.00     0.00    0.6031

Yearly
event      Median       Median                                  Pr(event|
rate:     survival,    survival,    Pr(event),    Pr(event),    no loss),
active     control      active       control       active       control

0.3700      0.750        1.500        0.8860        0.6737       0.8860

Pr(event|
no loss),    Hazard                 # events     # events     # events
 active      ratio     Log(HR)    (Ho-pooled)     (RGS)     (Freedman)
 0.6737      2.0000    0.6931          94           88           95

                              Total                                  Sample
               Total         Sample          Total       Sample      size
# events      Sample          size          Sample      size (no    (naive
  (L&F)     size (RGS)    (Freedman)    size (L&F)      loss)       loss)

   90           115           122           121          115         115
```

How do we pick from the feasible designs?

The first 4 designs all have 3 1/2 years total duration, since the follow-up period starts after the last patient has been accrued. The shorter the follow-up period given this fixed study duration, the more patients we have to enroll.

In some cases, it will be much more cost-effective to enroll fewer patients and follow them for longer. This corresponds to cases where the initial cost per patient is very high.

In other cases (where the initial cost per patient is lower), it will be better to enroll more patients. The median follow-up for the first 4 designs are 3, 2.5, 2.25, and 2 years, respectively. The total cost of treatment could be estimated by multiplying the number of patients by the median follow-up time.

I prefer to keep the accrual period as short as possible, given how many patients can feasibly be enrolled. This will tend to give the smallest number of patients among the feasible

designs. Which design would this correspond to?

Another issue to think about is whether the background conditions of the disease are changing rapidly (like AIDS) or are fairly stable (like many types of cancer). For the former situation, it would be best to have a study with a short duration so the results will have more interpretation.

Using the information given, there are a lot of other quantities we can calculate:

- The **hazard ratio of control to treated** is:

$$\frac{\text{median(Rx B)}}{\text{median(Rx A)}} = \frac{18}{9} = 2$$

- The **hazard rates** for the two treatment arms are:

$$\text{for Rx A:} \quad \lambda_0 = \frac{-\log(0.5)}{\text{median(Rx A)}} = \frac{-\log(0.5)}{9} = 0.0770$$

$$\text{for Rx B:} \quad \lambda_1 = \frac{-\log(0.5)}{\text{median(Rx B)}} = \frac{-\log(0.5)}{18} = 0.0385$$

- The **yearly probability of an event** is:

$$\begin{aligned} \text{for Rx A:} \quad Pr(T < 1|\lambda_0) &= 1 - e^{(-\lambda_0 * t)} \\ &= 1 - e^{(-0.0770*12)} = 0.603 \end{aligned}$$

$$\begin{aligned} \text{for Rx B:} \quad Pr(T < 1|\lambda_1) &= 1 - e^{(-\lambda_1 * t)} \\ &= 1 - e^{(-0.0385*12)} = 0.370 \end{aligned}$$

What would happen above if we used time t in years (i.e., t=1) instead of months?

What would happen if we calculated both the hazard rate and yearly event probability using time in years?

Based on a design with 2.5 years accrual and 1 year follow-up:

- The median follow-up time

$$\text{median FU} = A/2 + F = 30/2 + 12 = 27 \text{ months}$$

- The probability of an event during the entire study is: (using the approximation in notes)

$$\begin{aligned} \text{for Rx A:} \quad P_c &= 1 - exp(-\lambda_0 * [A/2 + F]) \\ &= 1 - exp(-0.0770 * 27) = 0.875 \\ \text{for Rx B:} \quad P_e &= 1 - exp(-\lambda_0 * [A/2 + F]) \\ &= 1 - exp(-0.0385 * 27) = 0.646 \end{aligned}$$

  The above numbers differ from what you'd get in the printout from my program, since I calculate the exact probability under the exponential distribution, instead of using the approximation)

In the calculations above, all of the "time" periods were in terms of months. You have to remember to keep the scale the same throughout.

To use my program, you need to translate the time scale in terms of years. So a median of 18 months survival would be entered as median=1.5.

What happens if we add loss to follow-up?

Required sample size for A=2.5, FU=1 year

Yearly Loss to Follow-up	Number of Events Required	Total Sample Size	Total Study Duration
0	88	122	3.5
5%	88	128	3.5
10%	88	133	3.5
20%	88	147	3.5

11.7 Sequential Design and Analysis of survival studies

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 can seriously inflate the true type I error of a trial. If interim analyses are to be performed, it is ESSENTIAL to carefully plan these in advance, and to adjust all tests appropriately so the the type I error is of the desired size.

How does the type I error become inflated?

Consider a two group study comparing treatments A and B.

Suppose the data are normally distributed (say $X_i \sim N(\mu_A, \sigma^2)$ in group A, and similarly for group B), so that the null hypothesis of interest is

$$H_0 : \mu_A = \mu_B$$

It is not too hard to figure out how the type I error can get inflated if a naive approach is used.

Suppose we plan to do K interim analyses, and that exactly m individuals will enter <u>each</u> treatment between each analysis. The test statistic at the kth analysis will be

$$Z_k = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(X_{Aij} - X_{Bij})/km}{\sqrt{2\sigma/km}} = \frac{\sum_{i=1}^{k} d_i/k}{\sqrt{2\sigma/km}}$$

where d_i is the difference between the two group means at the ith analysis,

$$d_i = \overline{X}_{Ai} - \overline{X}_{Bi}$$

and $\overline{X}_{Ai}$ and $\overline{X}_{Bi}$ are the means in groups A and B of the m individuals who entered in the i-th time period.

11.7.1 Naive Interim monitoring procedure:

- Allow m patients to enter on each treatment arm
  (total of $2m$ additional patients)
- Calculate Z_k based on the current data

- Reject the null hypothesis if $|Z_k| > z_{1-\alpha/2}$, where α is the desired type I error.

The overall type I error rate for the study is:

$$Pr(|Z_1| > z_{1-\alpha/2} \text{ or } |Z_2| > z_{1-\alpha/2} ... \text{ or } |Z_K| > z_{1-\alpha/2})$$

If the test at each interim analysis is performed at level α, then clearly this probability will exceed α. The table below shows the Type I error rate if each test is done at $\alpha = 0.05$ for various values of K:

Number of interim analyses (K)						
1	2	3	4	5	10	25
5%	8.3%	10.7%	12.6%	14.2%	19.3%	26.6%

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

For survival data, the calculations become MUCH more complicated since the data collected within each time interval continues to change as time goes on!

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

Pocock Approach

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

A problem with the Pocock method is that even the very last analysis has to be performed at level α'. 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.

11.7.2 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 analysis after each new individual enters the study!
  - There are also approaches where the randomization itself is modified as the trial proceeds. E.g. Zelen's "Play the winner rule" (New England Journal of Medicine 300, 1979, page 1242) and Ware's "ECMO" study (Statistical Science, 4, 1989, page 298)

- Some designs allow for early stopping in the absence of a sufficient treatment effect as the trial progresses. These procedures are referred to as "stochastic curtailment" or "conditional power" calculations.

- Designing a group sequential trial for survival data requires sophisticated and highly specialized software. EAST, a package from CYTEL SOFTWARE that does standard (fixed) survival designs, as well as sequential designs.

- Many "non-statistical" issues enter decisions about whether or not to stop a trial early

- P-values based on analyses of studies with sequential designs are difficult to interpret (see the Chapter by Emerson and Banks in *Case Studies in Biometry*)

- Once you do 5 interim analyses, then adding more makes little difference. The policy developed by SDAC (Statistical and Data Analysis Center for the ACTG) is to have randomized Phase III studies monitored at least once per year (for safety reasons), and most studies have 1-3 interim looks.

- going from a fixed to a group sequential design adds only about 3-4% to the required maximum sample size. This is a good rule of thumb to use in calculating the sample size when you plan on doing interim monitoring.