Max Planck Odense Center on the Biodemography of Aging

University of Southern Denmark

Power and sample-size determination

Virginia Zarulli, Adam Lenart

MaxO - SDU

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."
(R. Fisher)

Aim

Before starting a study, find out the sample size required to be able to discover a difference between survival curves or impact of a covariate for a given effect size, significance level and power.

- How many subjects are necessary to have enough power for a given hazard ratio?
- How large hazard ratio can we detect for a given sample size?

Type I and II errors

Null hypothesis: treatment has no effect.

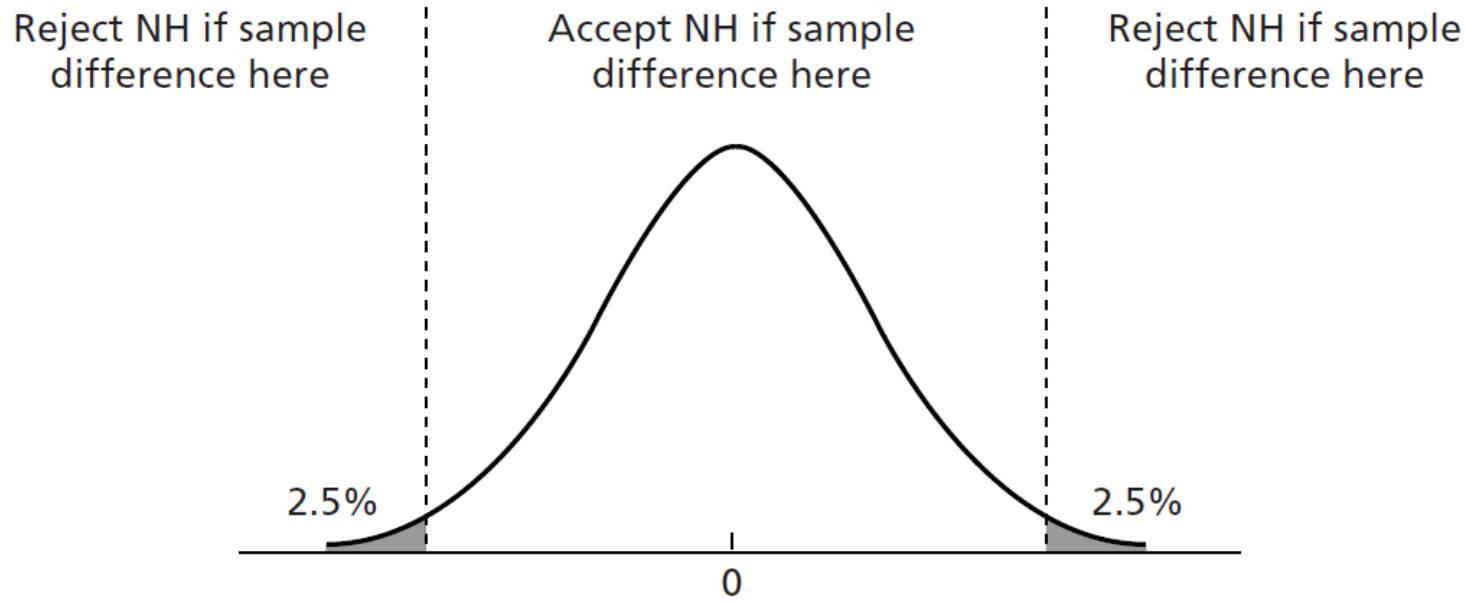
- **Type I error** Rejecting a true null hypothesis.
  - Treatment has no effect in reality but we reject this hypothesis and conclude that it does have an effect.
- **Type II error** Not rejecting a false null hypothesis.
  - Treatment has an effect in reality but we do not reject the null hypothesis and conclude that it does not have an effect.

Type I and II errors

	Null hypothesis	
	True	False
Reject	Type I error (probability = significance)	Correct (probability = power)
Not reject	Correct (probability = 1-significance)	Type II error (probability = 1-power)

- **Power**: correctly rejecting a false null hypothesis.

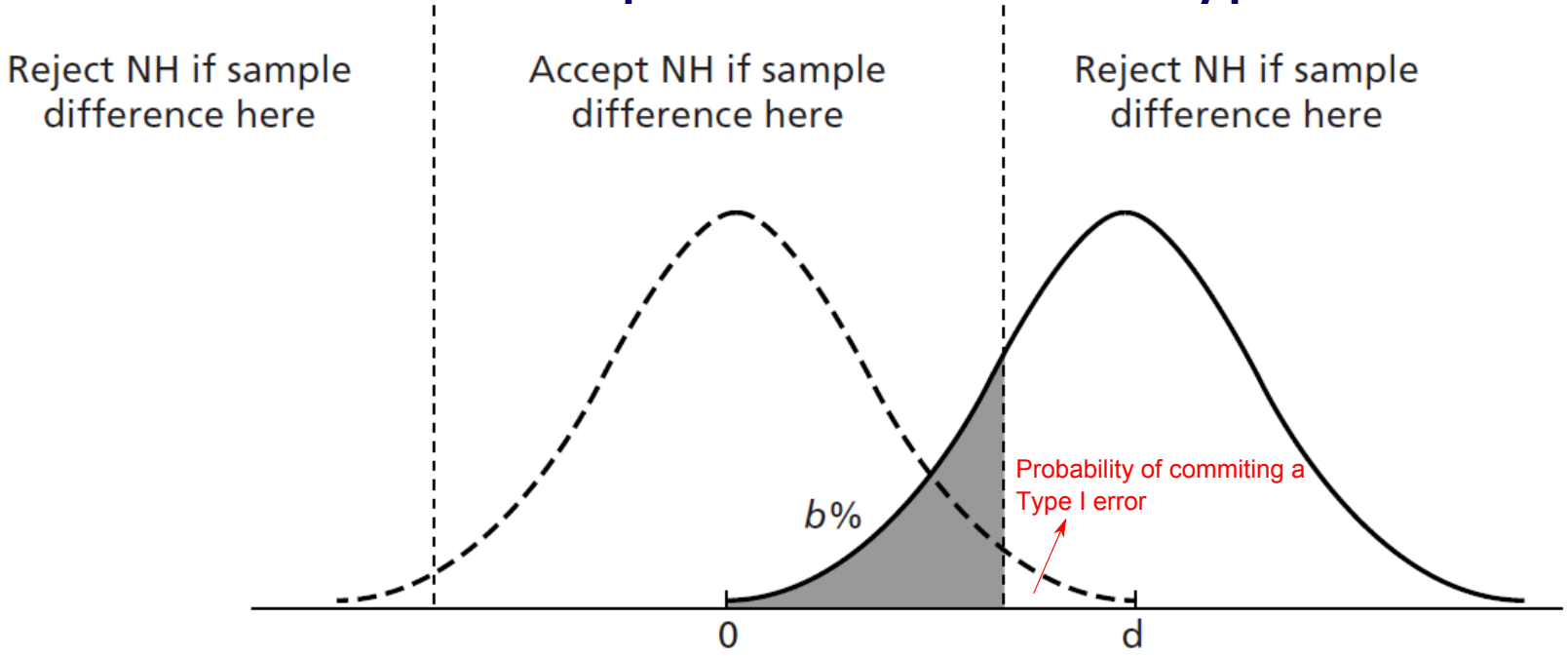
Type I error: reject a true null hypothesis



(a) Type I error. Null hypothesis (NH) is *true*. Population difference = 0. The curve shows the sampling distribution of the sample difference. The shaded areas (total 5%) give the probability that the null hypothesis is wrongly rejected.

(Kirkwood-Sterne 2003: Medical Statistics, 2nd ed. Blackwell. pp. 426.)

Type II error: do not reject a false null hypothesis



(b) Type II error. Null hypothesis is *false*. Population difference = d≠0. The continuous curve shows the real sampling distribution of the sample difference, while the dashed curve shows the sampling distribution under the null hypothesis. The shaded area is the probability (*b*%) that the null hypothesis fails to be rejected.

- Trade-off between committing a type I or type II error for a fixed sample size.

(Kirkwood-Sterne 2003: Medical Statistics, 2nd ed. Blackwell. pp. 426.)

Sample-size calculations in Stata

Stata offers sample-size or power calculations for detecting:

- non-parametric differences between two survival curves (log-rank test);
- difference between two exponential survival functions/hazard rates ;
- significant covariate in a Cox model.

Otherwise, for more complicated models or assumptions, we need to rely on simulations.

Minimal information

We need to provide information on at least,

- the required significance level (`alpha()`);
- the required power (`power()`);
- hazard ratio (`hratio()`);
- allocation of the number of subjects in control and treatment group (`p1()`).
- censoring (log-rank) and failure probability (Cox).

to be able to calculate the necessary sample size.

By default Stata uses $\alpha = 0.05$, power = 0.8, proportion of controls = 0.5 and hazard ratio = 0.5 and assumes no censoring.

Non-parametric difference between survival curves

```
stpower logrank s1, power(...)  alpha(...)
          p1(...)  hratio(...)  schoenfeld
```

*s*1 is a number that stands for the proportion of censored in the control group.

Using the Schoenfeld method, the estimated number of events,

$$E = \frac{\left(z_{1-\alpha/k} + z_{1-\beta}\right)^2}{p_1(1-p_1)\ln^2\left(\frac{h_2}{h_1}\right)},$$

where z_n denotes the nth percentile of the standard normal distribution, k=2 if the test is two-sided (default) or 1 if it is one-sided, p_1 is the proportion of subjects in the control group and h_2/h_1 stands for the ratio of treatment to control hazard.

Example: non-parametric difference between survival curves

Clinical trial to assess new treatment for patients with chronic active hepatitis. Under standard treatment, 41% of patients survive beyond 5 years. We expect new treatment to increase survival beyond 5 years to 60% (Collett: *Modelling Survival Data in Medical Research*, 10.1)

We would like to have a significant result 4 out 5 times at a significance level of 0.05. We will enroll an equal number of treatment and control subjects.

Example: non-parametric difference between survival curves contd.

We can calculate the hazard ratio (Δ) from

$$S_T(t) = [S_C(t)]^{\Delta}$$
$$0.6 = 0.41^{\Delta}$$
$$\Delta = \frac{\log(0.6)}{\log(0.41)} \approx 0.57$$

assuming proportional hazards to use

```
stpower logrank 0.41, hratio(0.57) power(0.8)
alpha(0.05) p1(0.5).
```

We could have also used

```
stpower logrank 0.41 0.6, power(0.8)
alpha(0.05) p1(0.5)
```

to get the same result.

Example: non-parametric difference between survival curves contd.

```
. stpower logrank 0.41, hratio(0.57) power(0.8) alpha(0.05) p1(0.5)

Estimated sample sizes for two-sample comparison of survivor functions
Log-rank test, Freedman method
Ho: S1(t) = S2(t)

Input parameters:

      alpha =    0.0500  (two sided)
         s1 =    0.4100
         s2 =    0.6016
     hratio =    0.5700
      power =    0.8000
         p1 =    0.5000

Estimated number of events and sample sizes:

          E =       106
          N =       212
         N1 =       106
         N2 =       106
```

number of controls

number of treated

If we are certain that treatment is better (worse) than control, then we can add the `onesided` option to reduce the required sample size.

Example: non-parametric difference between survival curves contd.

We can even be a little bit smarter with our resources by enrolling more patients to the treatment arm.

```
. stpower logrank 0.41, hratio(0.57) alpha(0.05) p1(0.2(0.1)0.6) n(212)

Estimated power for two-sample comparison of survivor functions
Log-rank test, Freedman method
Ho: S1(t) = S2(t)
```

Power	N	N1	N2	E	S1	S2	HR	Alpha*
.713272	212	42	170	93	.41	.601571	.57	.05
.791451	212	63	149	97	.41	.601571	.57	.05
.813041	212	84	128	101	.41	.601571	.57	.05
.800525	212	106	106	105	.41	.601571	.57	.05
.756221	212	127	85	109	.41	.601571	.57	.05

* two sided

Difference between two exponential survival functions

`stpower exponential` *h1*`, power(...) alpha(...) p1(...) hratio(...)`

*h*1 is the baseline hazard for the control group.

- by assuming exponential distribution, we can estimate sample size and power for more flexible study designs.
  - censoring: instead of survival probability at the end of study (log-rank test), we can specify the end of the study time, `fperiod()`.
  - attrition: losses to follow up, we can specify a hazard rate for losing subjects in the control and treatment groups by `losshaz(`h_{l1}`,` h_{l2}`)`.

Example: Difference between two exponential survival functions

Survival under current treatment after one year is 50%. A new treatment proposes to increase one year survival to 72%. The study will continue for three years, 10% of patients are estimated to be lost each year in both the control and treatment arms.
We would like to have the power of the trial to be 90% and have a significance level of 0.05.

Example: Difference between two exponential survival functions contd.

From $S(t) = \exp(-\lambda t)$, we can calculate the hazard rates (h) for both survival and attrition (ω) as

$$\lambda t = -\log(S(t)),$$

to get $h_1 \approx 0.7,\ h2 \approx 0.33,\ \omega \approx 0.105$

Example: Difference between two exponential survival functions contd.

```
. stpower exponential 0.7, hratio(0.47) power(0.9) alpha(0.05) fperiod(3) losshaz(0.105 0.105)
Note: input parameters are hazard rates.

Estimated sample sizes for two-sample comparison of survivor functions
Exponential test, hazard difference, conditional
Ho: h2-h1 = 0

Input parameters:

       alpha =    0.0500  (two sided)
          h1 =    0.7000
          h2 =    0.3290
       h2-h1 =   -0.3710
       power =    0.9000
          p1 =    0.5000

 Accrual and follow-up information:

    duration =    3.0000
   follow-up =    3.0000
         lh1 =    0.1050
         lh2 =    0.1050

Estimated sample sizes:

           N =       120
          N1 =        60
          N2 =        60
```

Detecting effect of a covariate in a Cox model

```
stpower cox, hratio(...)  power(...)
alpha(...)  sd(...)  r2(...)  failprob(...)
```

where `sd()` and `r2()` are the standard deviation of analyzed covariate and the multiple correlation coefficient (squared) of the covariate with the other covariates in the model. `failprob()` denotes the probability to fail (1- probability of being censored).

Detecting effect of a covariate in a Cox model

$$E = \frac{\left(z_{1-\alpha/k} + z_{1-\beta}\right)^2}{\sigma^2 \beta_1^2 (1 - R^2)},$$

where the estimated number of events depend on the standard normal distribution, the standard deviation (σ), the hazard ratio (β_1) of covariate X_1 and its squared multiple correlation coefficient with the other covariates in the model.

Example: Detecting effect of a covariate in a Cox model

We would like to investigate the association between blood urea nitrogen level (*lnbun*) of multiple myeloma patients and death.

We predict that a higher level of nitrogen is an indicator of worse renal function and is positively correlated with death.

We are interested in finding out whether a unit increase in *lnbun* leads to a 50% higher risk of death adjusting for other covariates in our data.

From a pilot study, we guess the standard deviation and the multiple correlation coefficient that are necessary to calculate the sample size in a Cox model. In the pilot study, we also saw that 48 people died out of 65, giving us a failure probability of $48/65 = 0.738$.

Example: Detecting effect of a covariate in a Cox model contd.

```
. use Q:\Teaching\SurvivalAnalysis\slides\Day4\data\myeloma.dta
(Multiple myeloma patients)

. summarize lnbun
```

Variable	Obs	Mean	Std. Dev.	Min	Max
lnbun	65	1.3929	.3126297	.7782	2.2355

```
. regress lnbun hemo platelet age lnwbc fracture lnbm protein scalcium
```

Source	SS	df	MS
Model	1.15026278	8	.143782848
Residual	5.10492732	56	.091159416
Total	6.2551901	64	.097737345

Number of obs	65
F(8, 56)	1.58
Prob > F	0.1525
R-squared	0.1839
Adj R-squared	0.0673
Root MSE	.30193

lnbun	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hemo	-.0043198	.0182425	-0.24	0.814	-.0408639	.0322242
platelet	-.0468442	.1301531	-0.36	0.720	-.3075722	.2138838
age	.0075177	.0041724	1.80	0.077	-.0008405	.015876
lnwbc	.429355	.1675451	2.56	0.013	.0937219	.7649882
fracture	-.0874605	.0919659	-0.95	0.346	-.2716904	.0967694

Example: Detecting effect of a covariate in a Cox model contd.

```
. stpower cox, failprob(0.738) power(0.9) alpha(0.05) hratio(1.5) sd(0.313) r2(0.184) onesided

Estimated sample size for Cox PH regression
Wald test, log-hazard metric
Ho: [b1, b2, ..., bp] = [0, b2, ..., bp]

Input parameters:

         alpha =    0.0500  (one sided)
            b1 =    0.4055
            sd =    0.3130
         power =    0.9000
     Pr(event) =    0.7380
            R2 =    0.1840

Estimated number of events and sample size:

             E =       652
             N =       883
```

Accrual time

Previously, we assumed that we have all of the patients ready for the study instantaneously at time 0. However, in practice, we might have an accrual period, R, during which subjects are recruited (and followed) and a follow-up period during which the subjects are only followed, f.

$$T = R + f$$

- Log-rank: uniform accrual
  - same number of subjects recruited at each time unit
- Exponential: uniform and exponential accrual
  - exponential accrual assumes a right-truncated exponential distribution.
- Cox: no accrual, we can decrease `failprob()` to account somewhat for it.

Accrual time in log-rank models

When there is accrual, we need to calculate the probability of failure conditionally on the entry times.

$$p_E = 1 - \frac{1}{R}\int_f^T \tilde{S}(t)\,dt$$

for $\tilde{S}(t) = p_1 S_1(t) + (1-p_1)S_2(t)$ By Simpson's rule

$$p_E \approx 1 - \frac{1}{6}\left[\tilde{S}(f) + 4\tilde{S}\left(\frac{R}{2}+f\right) + \tilde{S}(T)\right]$$

Example: accrual time for a log-rank test

Clinical trial to assess new treatment for patients with chronic active hepatitis. Under standard treatment, 41% of patients survive beyond 5 years. We expect new treatment to increase survival beyond 5 years to 60%. We enroll patients for 2 years and have a follow-up period of 3 years. We expect to lose 15% of the patients (`wdprob(0.15)`).
In the control group, we expect to have a survival of 70% of the patients after 3 years and 52% of the patients after 4 years.

Example: accrual time for a log-rank test contd.

```
. stpower logrank, hratio(0.57) power(0.8) alpha(0.05) p1(0.5) simpson(0.7 0.52 0.41) wdprob(0.15)
Note: probability of an event is computed using Simpson's rule with
      S1(f) = 0.70, S1(f+R/2) = 0.52, S1(T) = 0.41
      S2(f) = 0.82, S2(f+R/2) = 0.69, S2(T) = 0.60

Estimated sample sizes for two-sample comparison of survivor functions
Log-rank test, Freedman method
Ho: S1(t) = S2(t)

Input parameters:

      alpha =    0.0500  (two sided)
     hratio =    0.5700
      power =    0.8000
         p1 =    0.5000
 withdrawal =     15.00%

Estimated number of events and sample sizes:

          E =       106
          N =       320
         N1 =       160
         N2 =       160
```

Accrual time in exponential model

- **Uniform accrual**: same number of subjects entering the study at each time period until the specified sample size is reached (`aperiod()`).
- **Truncated exponential accrual**: subjects enter the study in an exponentially increasing/decreasing number at each time period until sample size is reached according to:

$$G(t) = \frac{1 - \exp(-\lambda t)}{1 - \exp(-\lambda R)} \quad 0 \leq t \leq R$$

if $\lambda < 0$, accrual starts slowly and increases exponentially and when $\lambda > 0$, accrual starts fast and then slows down, $\lambda = 0$ gives uniform accrual.

Example: Accrual time in exponential model

Survival under current treatment after one year is 50%. A new treatment proposes to increase one year survival to 72%. The study will continue for three years, 10% of patients are estimated to be lost each year in both the control and treatment arms. The accrual period is one year, we expect to be able to enroll patients faster in the beginning of the trial and somewhat slower later.
We would like to have the power of the trial to be 90% and have a significance level of 0.05.

Example: Accrual time in exponential model contd.

```
. stpower exponential 0.7, hratio(0.47) power(0.9) alpha(0.05) fperiod(2) losshaz(0.105 0.105) aperiod(1) ashape(0.2)
Note: input parameters are hazard rates.

Estimated sample sizes for two-sample comparison of survivor functions
Exponential test, hazard difference, conditional
Ho: h2-h1 = 0

Input parameters:

       alpha =    0.0500  (two sided)
          h1 =    0.7000
          h2 =    0.3290
       h2-h1 =   -0.3710
       power =    0.9000
          p1 =    0.5000

 Accrual and follow-up information:

    duration =    3.0000
   follow-up =    2.0000
     accrual =    1.0000  (exponential)
  accrued(%) =     50.00  (by time t*)
          t* =    0.4750  (47.50% of accrual)
         lh1 =    0.1050
         lh2 =    0.1050

Estimated sample sizes:

           N =       128
          N1 =        64
          N2 =        64
```