KM－HR

The Product-limit method (= Kaplan-Meier method)

- 生存機能はS(t) で表され、ある個人が時間(t)を超えて生存する確率を示している。生存機能S(t) を生存・死亡に限る必要はなく、エンドポイントは再発、退院、離婚、何でもよい。例えばS(1095)は、「遠隔転移のない食道癌患者において、手術後1000日を経た時点で再発なく生存している確率は59%である」といった具合である（図1）。

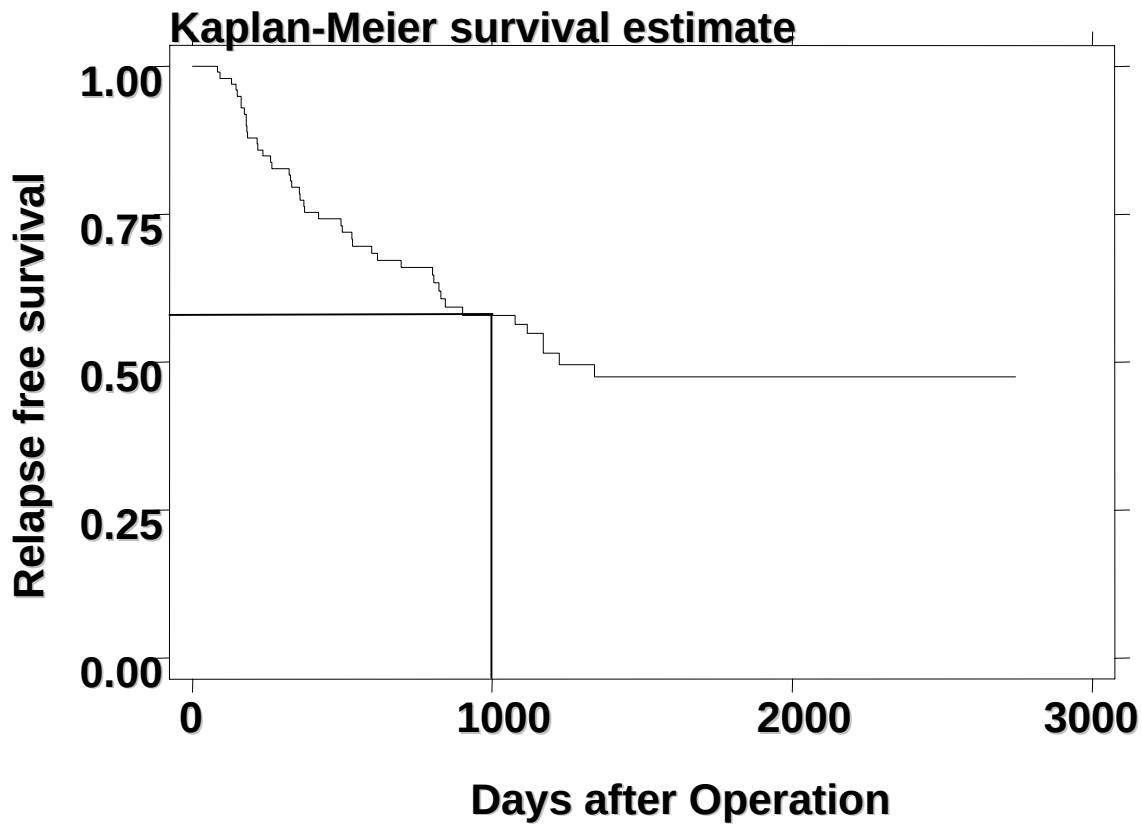


図1. 食道癌患者手術症例のKaplan-Meier 生存曲線

- 逆に、半数など一定の割合の患者が生存できる期間で表現することもある。例えば図2はT2以上の手術病理標本でエンドセリン免疫染色陽性と非陽性で再発なく生存している期間を比較した。前者では半数が1年以内に再発しているのに対して、後者では半数以上が3年以上再発無く生存している。

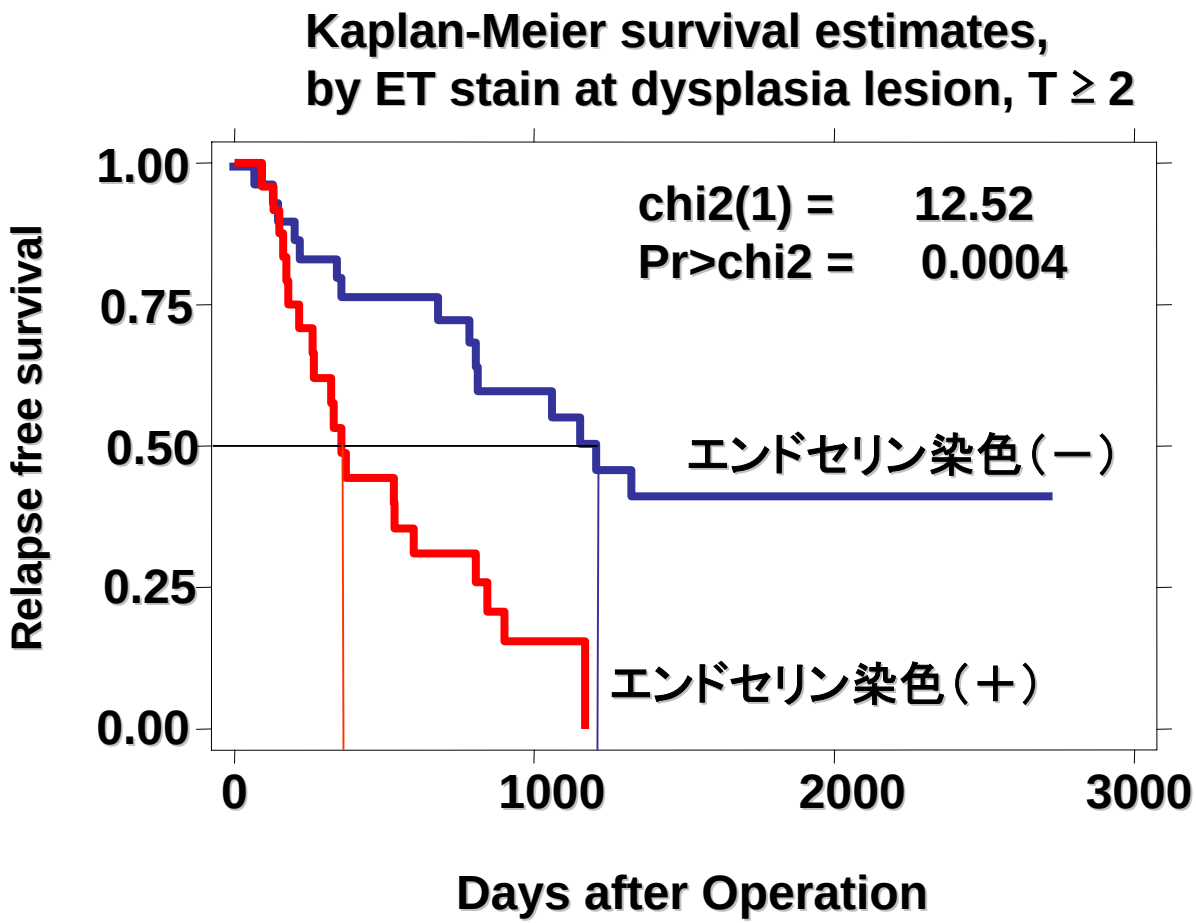


図2. 食道癌患者手術症例（T1、M1を除く）のKaplan-Meier 生存曲線
病理切片免疫染色におけるエンドセリン染色例、非染色例

- 今から50年以上前、シドニー・ファーバー博士らがアミノプテリンを使用して、小児白血病を高率で寛解導入に成功した。しかし、多くが中枢神経を含む箇所に再発した。ここに、12人の白血病患児がいるとする。全員が寛解に入っているが、やがて再発（エンドポイント）するとする。これを1-2ヶ月の間に再発した患児数は1、2－3ヶ月の場合は1, といった具合に算定すると、生命表の原理となってしまう。そこで患児1人が再発する毎にS(t)を計算することにする。よって、再発までの期間はまちまちである（表1）。

表1．小児急性リンパ性白血病(ALL)第一寛解期にある12人の再発までの期間

患者番号	1	2	3	4	5	6	7	8	9	10	11	12
再発までの期間（月）	2	3	6	6	7	10	15	15	16	27	30	32

- 最初の時点では12人全員が寛解状態にあるから、S(0) = 1である。しかし2ヶ月の時点で1人再発している。さてS(2)は2ヶ月を超えて寛解を維持する確率であるが、2ヶ月以内に再発する確率を1から引いた値ともいえる。2ヶ月の時点で12人中1人が再発しているということは、2ヶ月までに再発する確率は1/12 = 0.0833と考えられる。よって2ヶ月を超えて再発せずに寛解を維持する確率は、1 –0.0833 = 0.9167 となる。S(0) = 1で、S(2) = S(0) x (1 – 0.0833) = 1 x 0.9167 = 0.9167となる。さて次に3ヶ月の時点でもう1人再発している。2ヶ月から3ヶ月の間に1人が再発する確率は1/11 = 0.0909である。既に2ヶ月までに1人再発しており、更に3ヶ月までに1人再発したわけであるから、3ヶ月を超えて生存する確率はS(3) = S(2) x (1 – 0.0909) = 0.9167 x 0.9091 = 0.8334のようになる。

表2. Kaplan-Meier 生存曲線の手計算

月	q_t*1	1 - q_t*2	L_t*3	S(t)
0	0.0000	1.0000	12	1.0000
2	0.0833	0.9167	11	0.9167
3	0.0909	0.9091	10	0.8333
6	0.2000	0.8000	8	0.6667
7	0.1250	0.8750	7	0.5833
10	0.1429	0.8571	6	0.5000
15	0.3333	0.6667	4	0.3333
16	0.2500	0.7500	3	0.2500
27	0.3333	0.6667	2	0.1667
30	0.5000	0.5000	1	0.0833
32	1.0000	0.0000	0	0.0000

*1: qt は時間tまでに再発する確率。*2: 1 - qtは、再発せずに時間tを超える確率。
*3:lt は時間tの時点で寛解にある人数

- このように手計算を続けていったのが表２である。これを生存曲線として描くと図３のようになる。

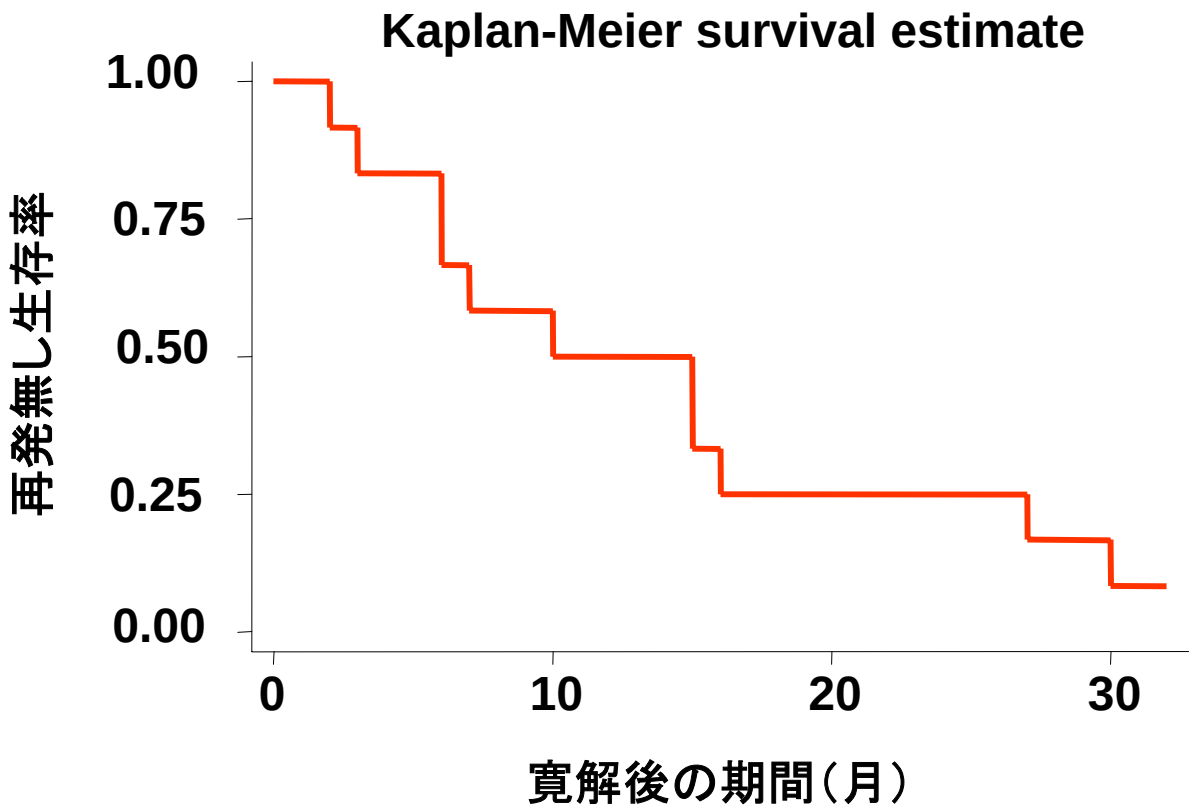


図3. 小児白血病患者の再発までのKaplan-Meier 生存曲線

センサー

- 例えば骨髄移植後の白血病患者さん100人の再発について経過観察しているとする。しかし、観察期間も5年になり、予定通り研究を打ち切ろうと思う。しかし全員が5年前から経過観察しているわけではなく、2年前、1年前、あるいはつい先月から経過観察し始めた人もいる。現実的に生存を永遠に追跡することはできないので、いつかは中止しなくてはならない。よって通常経過途中の患者さんがでてくる。治療後4年経っている患者さんの再発は少ないかもしれないが、治療後1ヶ月の人は今後再発するかどうか全くわからない。それではこれらのデータを捨ててしまうのか。それはもったいない話である。それではこれらをセンサーという特別な扱いでデータに残すことにしよう。センサーには経過途中で観察が中止になった場合に加え、外来に来なくなってしまって観察できない場合、またもし骨髄移植後拒絶反応で死亡してしまった場合などもセンサーと考えることができる。しかし、少なくとも死亡するまでは再発していなかったとすれば、それなりのウエイトを置いてデータに組み込むことは可能である。

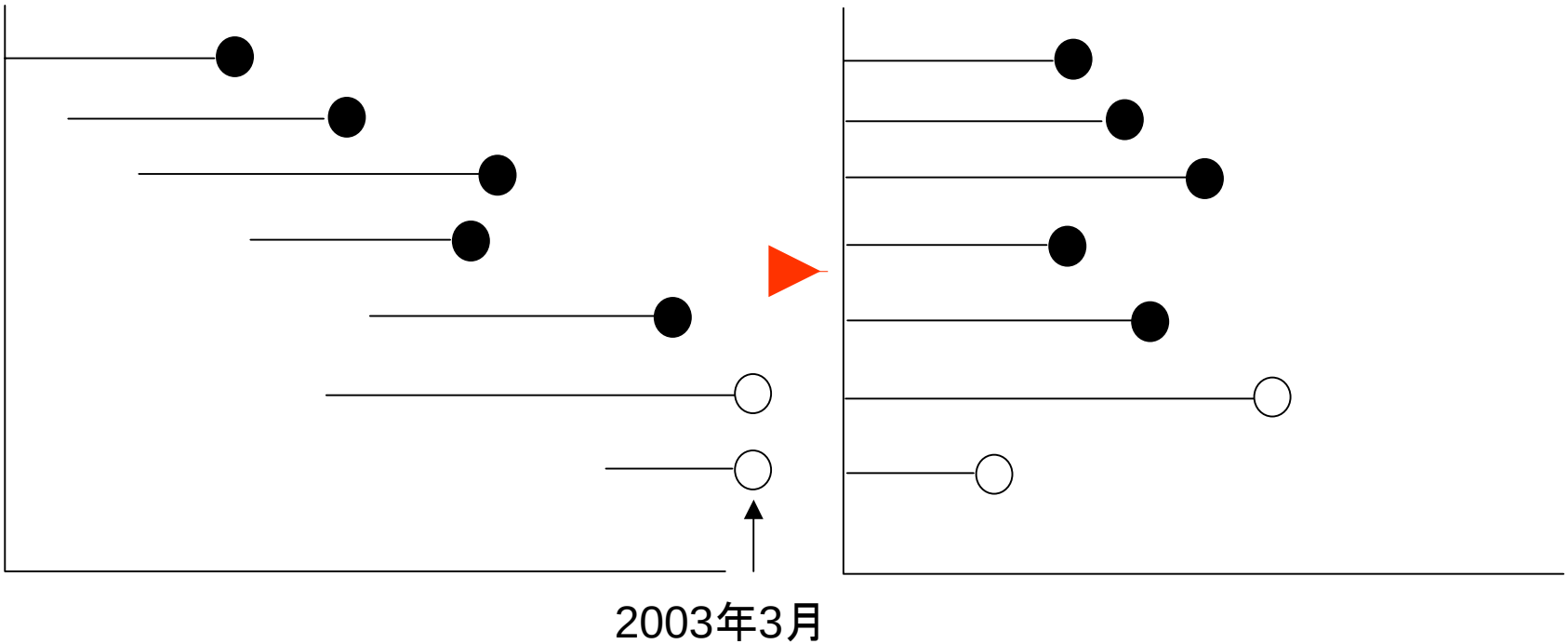


図4. 例えば治療後7人の患者さんをフォローしていたとする。
5人は残念ながら途中で亡くなった。2003年3月で経過観察を理由があって
打ち切ったとする。そうすると2人はセンサーとして数えられる。
上の図を整理すると下のようになる。

- 先の例で2番目と6番目の患児がそれぞれ3ヶ月、10ヶ月の時点で調査が修了しセンサー となったとする。そのような場合、慣例として数値の脇に＋を付してセンサー であることを示す（表３）。

表３　小児ALL第一寛解期12人の再発までの期間

患者番号	1	2	3	4	5	6	7	8	9	10	11	12
再発までの期間（月）	2	3+	6	6	7	10+	15	15	16	27	30	32

症例2と症例6がそれぞれ3ヶ月と10ヶ月でセンサーとなった場合

- 先と同様な操作を行ってみる。3ヶ月と10ヶ月では再発がないわけであるから、S(t)は1つ前のものと同じになる点に注意しなければならない（表4）。例えば、6ヶ月の時点では11人でなく10人中2人が再発したと考え、q6は0.2となる。しかしS(3) はセンサー でない状態では1人再発しているので0.8333だったが、センサー の場合再発していないのでS(3) はS(2) と同じになり、S(6)は0.9167に(1 – 0.2) をかけてやることになり、センサーのない場合の0.6667より少し生存曲線が改善された状態0.7333となる。これは、前の例では3ヶ月の時点で患児が再発したのに対し、今回は患児が少なくとも3ヶ月以上生存しているわけだから、当然の結果として生存曲線は少し改善することになる。

表4. センサーを加味したKaplan-Meier 生存曲線の手計算

月	q_t	1 - q_t	L_t*3	S(t)
0	0.0000	1.0000	12	1.0000
2	0.0833	0.9167	11	0.9167
3	0.0000	1.0000	11	0.9167
6	0.2000	0.8000	9	0.7333
7	0.1250	0.8750	8	0.6417
10	0.0000	1.0000	8	0.6417
15	0.3333	0.6667	6	0.4278
16	0.2500	0.7500	5	0.3208
27	0.3333	0.6667	4	0.2139
30	0.5000	0.5000	3	0.1069
32	1.0000	0.0000	2	0.0000

*1: qt は時間tまでに再発する確率。*2: 1 - qtは、再発せずに時間tを超える確率。
*3:lt は時間tの時点で寛解にある人数

- これを生存曲線として描くと図４のようになる。
- センサーとなった患児に関して、観察期間のところに印をつけるのが慣例である。
- グラフ中の標はセンサー を示している。3ヶ月と10ヶ月で1人ずつがセンサー となっている。

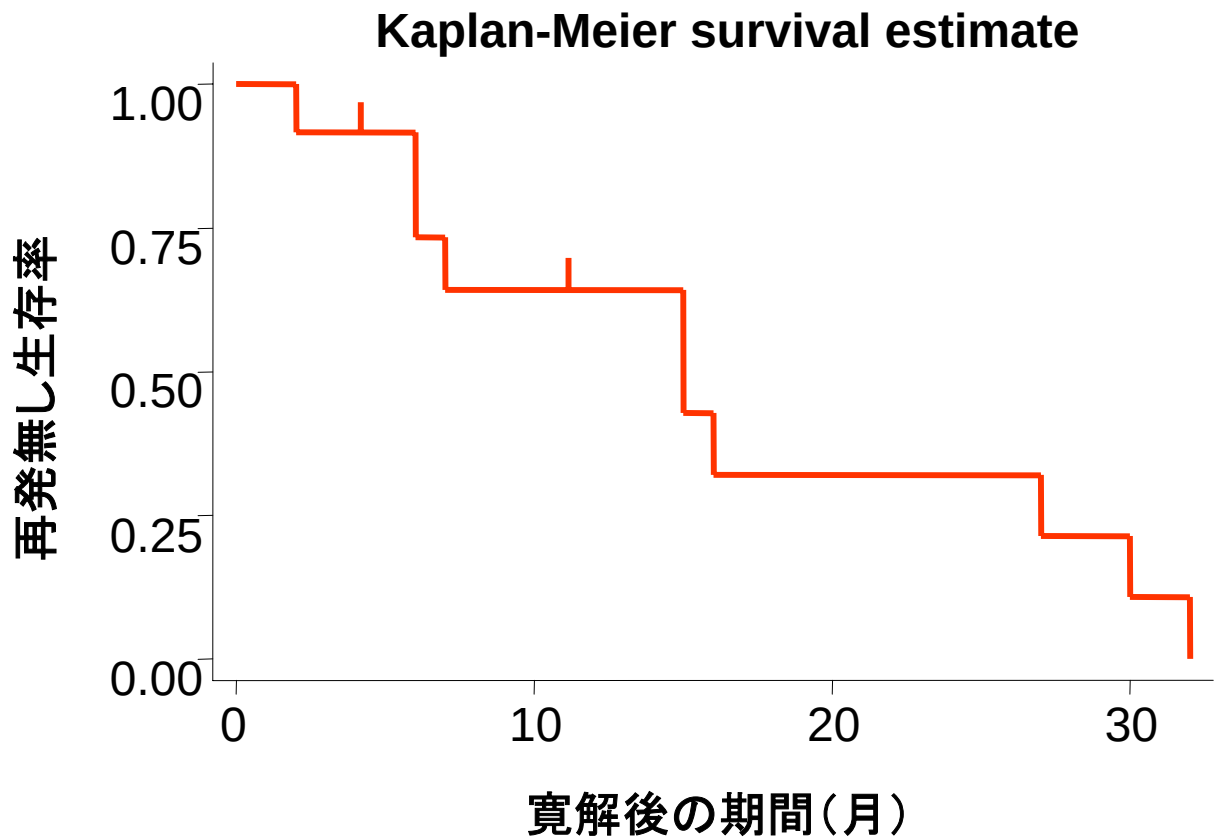


図5. センサーを加味した小児白血病患者の再発までのKaplan-Meier 生存曲線

生存曲線の比較検定：Log-rank test

- 　カプランマイヤー生存曲線において2つ以上の曲線を比較する際用いられる検定がlog rank test である。原理はカイ二乗検定で、2つの生存曲線が同じである場合の期待される値(expected)をだし、それと観察された値(observed)がどれくらい隔たっているかを計算し、極端に隔たっていれば、統計学的に有意差ありと結論するのである。t検定で2つは同じであるという帰無仮説から入ったのと似ている。
- AとBの治療があり、AとBの治療効果は等しく、よって生存曲線も同じになると仮定する。例えば図6のような状況のとき、単純に考えれば、治療Aで3人死亡がでているからAの方が悪いかもしれない。

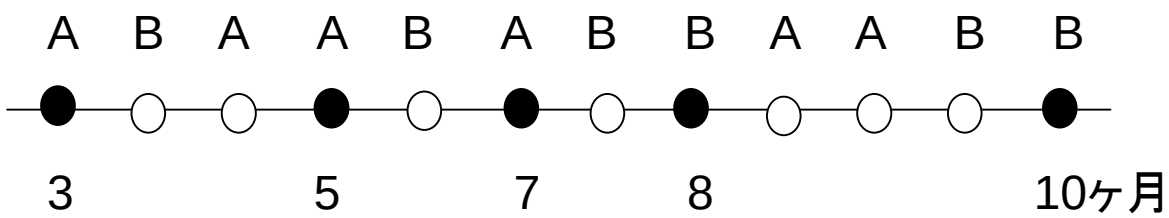


図6. 例えば、3ヶ月、5ヶ月、7ヶ月の時点でAの治療群から死亡が1人ずつ、一方8ヶ月、10ヶ月の時点でBの治療から死亡がでているとする。その他の患者さんは治療経過観察が途中で打ち切られた為センサー*となった。

- しかも、早期に死亡がでている。実際のところはどうなのだろうか？
- それぞれ死亡者が出た時点で、2 x 2 table を作る（表5）。まずは3ヶ月の時点である。常にAを軸にして考えるようにしよう。12人の患者さんはA、B 6人ずつに振り分けられている。そしてAに1人の死亡を観察した。ここでAとBの治療は同じであると仮定しているから、ここでAが死亡する確率は6/12 = 0.5 であることが期待される（同じ確率でBが死亡していたかもしれない）。
- (observed) – (expected) = (1 – 6/12) = 0.5
- そして、variance は
- var = (1 x 11 x 6 x 6)/{12 x 12 x (12-1)} = 0.25

- 同様に5ヶ月の時点ではどうだろうか？5ヶ月を超えてセンサーになったものは5ヶ月の時点で生存しているから含める。一方5ヶ月未満でセンサーになったものは含めない。そしてAに1人の死亡を観察した。上の図を見ながら数値を入れると上の表のようになる。ここでAとBの治療は同じであると仮定しているから、Aが死亡する確率は単純に9個から4つ選んだときAを選ぶ確率と同義であり、4/9 = 0.44 であることが期待される。
- (observed) – (expected) = (1 – 0.44) = 0.56
- var = (1 x 8 x 4 x 5)/{9 x 9 x (9-1)} = 0.25

- 同様に8ヶ月の時点ではどうだろうか？Aを軸に考えているから観察された死亡はOである。ここでAとBの治療は同じであると仮定しているから、ここでAが死亡する確率は2/5 = 0.4 であることが期待される。
- (observed) – (expected) = (0 – 0.4) = 0.6
- var = (1 x 4 x 2 x 3)/{5 x 5 x (5-1)} = 0.24
- 最後に全ての2 x 2 table を合わせる。
- [(1 – 6/12) + (1 – 4/9) + (1 – 3/7) + (0 – 2/5) + (0 – 1/1) – 0.5]2= 0.073
- 0.25 + 0.25 + 0.23 + 0.24 + 0.25 = 1.22
- 0.073 / 1.22 = 0.06 < 3.84 (c21)
- よって治療A、Bの間には差がないと結論できる。

表5

3ヶ月時点のまとめ

	死亡	生存	合計
A	1	5	6
B	0	6	6
合計	1	11	12

5ヶ月時点のまとめ

	死亡	生存	合計
A	1	3	4
B	0	5	5
合計	1	8	9

8ヶ月時点のまとめ

	死亡	生存	合計
A	0	2	2
B	1	2	3
合計	1	4	5

Hazard Function

- 生存機能はS(t) で表わされ、ある人が時間(t)を超えて生存する確率を示している。たとえば生命保険会社のセールスマンの「あなたは現在41歳ですが、65歳を超えて生存する確率は90%と予想されます」といった表現に代表される。一方、ハザード機能では、例えば「白血病に対しての骨髄移植では、移植後1年以内に再発することが多いが、その後再発はほとんど認められない」といった具合である。つまり時々刻々と変化するエンドポイント発生のリスクをとらえようとしているのだ。Kaplan-Meier生存曲線は半年を越えて生存する確率といった広い視野にたったものであり、Hazardカーブは半年を超えた人が次の１週間生存する確率といった、もっと短い期間に焦点を当て考えている。時々KM生存曲線をひっくり返したものがHazard function curve だと誤解している人がいるが、全然違う。

- 例えばあなたが‘東名高速道路を時速100kmで自動車を運転している状況を連想すると理解しやすいかもしれない（図6）。東京まで100kmだとすると、このペースで走れば1時間後に東京に着くはずである。でも川崎あたりから渋滞しているかもしれない。つまり最初の予想はpotential であり、今の状態が続けば(at a given moment)’ということになる。であるからこれから1時間以内に渋滞にはまる可能性はhazard function で示される。次のある期間（時間）内に事象が起こる率をh(t)で示す。しばしばHazard function h(t) は、ある人がある期間の最初まで生存し、その短い期間中に死亡（事象発生）する条件付き確率として用いられる。

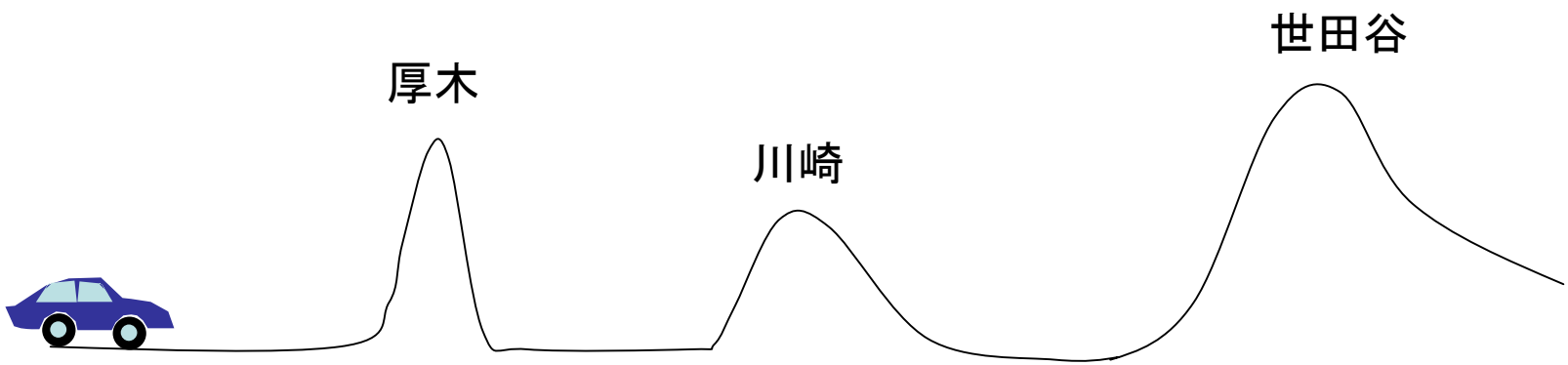


図6. 渋滞ハザードモデル。刻々と変化する渋滞リスクを示す。

- 臨床例でみてみよう（図7）。非常にリスクの高い手術では術後ICUで亡くなる患者さんが多いかもしれないし、吸入炭疽菌患者は入院後4日以内の死亡が多いが、7日以上もてば死亡することはまずない、癌患者は診断確定後すぐに亡くなることは無いが、再発や転移が増えていき、死亡するものが徐々に増えていく、しかし5年を過ぎると再発は少なくなる、云々といった具合である。

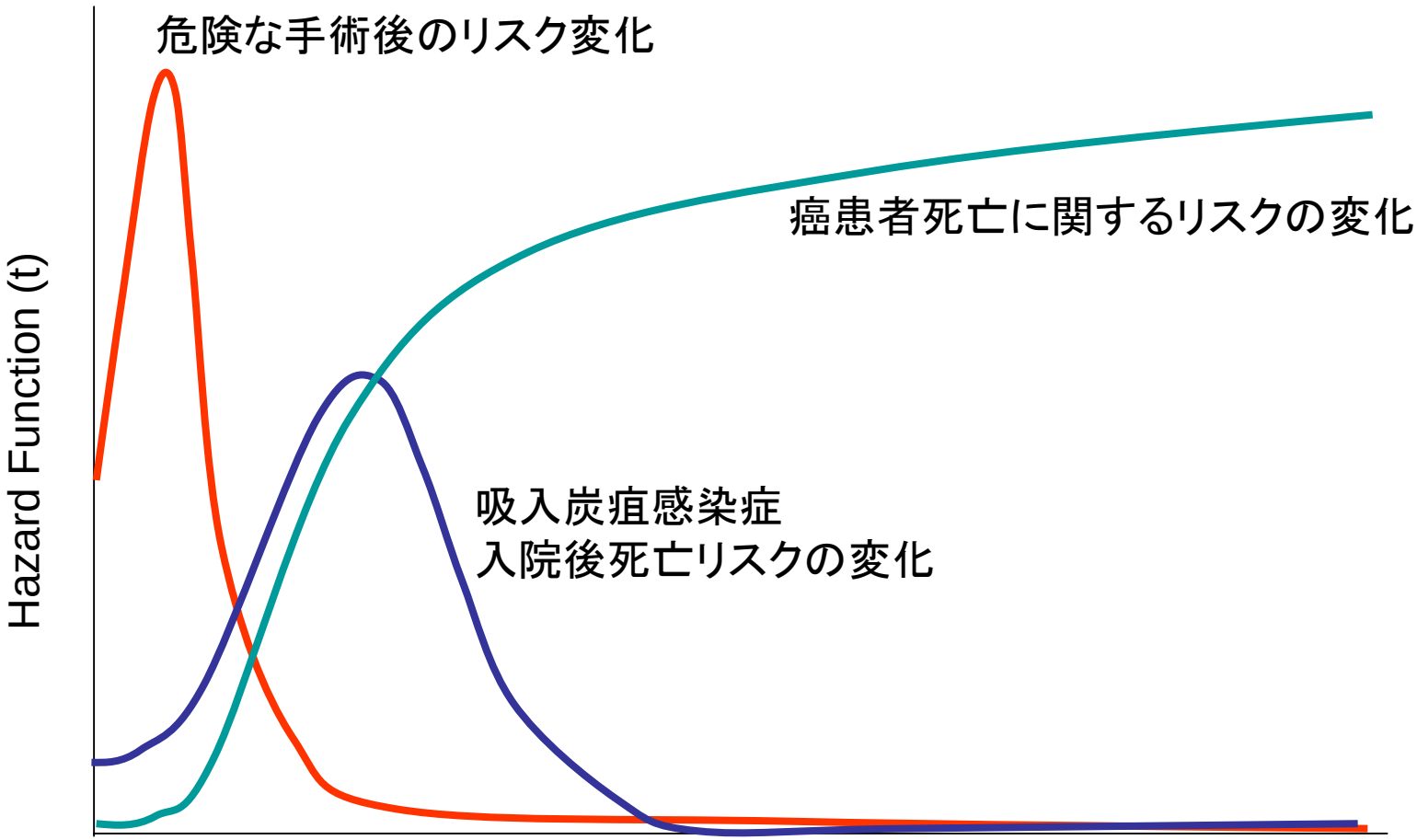


図7. 渋滞ハザードモデル。刻々と変化する渋滞リスクを示す。

Cox Hazard Model

- KMカーブを比較するのにLog rank test を用いたようにHazard curve の比較ではCox's model を用いる。例えばAとBの治療のHazard curve を比較するとする。Mまずは2つのHazard Curve が同じであると仮定する。そして反対として、2つの曲線は同じ形をしているが、constant k 倍ずれていると考える（図9）。
-
- H_0: $\lambda_A(t) = \lambda_B(t)$
- H_A: $\lambda_A(t) = \kappa\, \lambda_B(t)$　　κ = constant
-
- Proportional hazard: $\lambda_A(t) / \lambda_B(t) = \kappa$
-
- このように説明すると難しいのであるが、BはAより2倍良いということを統計学的に証明すると考えると理解しやすい。

生存曲線における対象数の計算

- 小児癌Xに対す現在の治療では患者の50%生存期間は1年である。新しい治療法により50%生存期間を1.5倍にまで延長することを期待できるとすると、ランダム化臨床試験を行う際、何人の対象数が必要になるか。

  ただし、α error = 0.05, power = 0.80 とする。

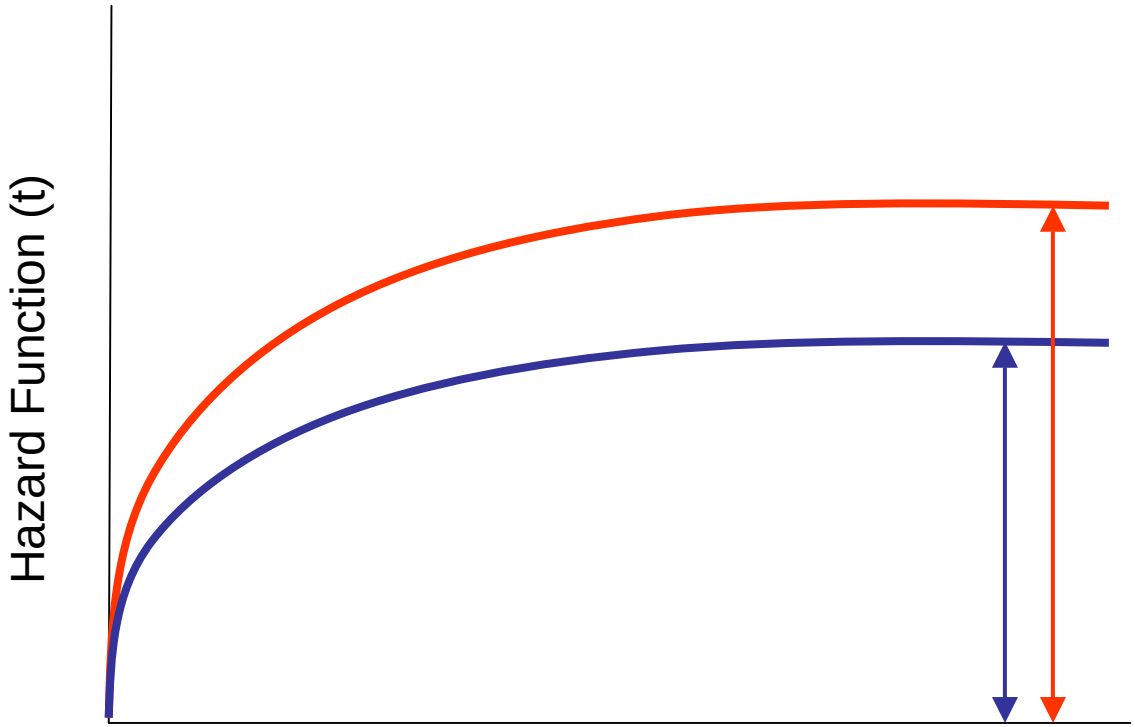


図9. 小児癌Xの標準治療と新規治療の
hazard curve がk倍異なると想定する。
両曲線が同じような形であればproportional
hazard model で比較できるが、形が異なる
場合にはnon-proportional hazard model と
なる。

今までの予後因子を上回れるか？

- Kaplan-Meier 生存曲線は、1つの因子についてしか検討比較できない。1つの因子とはいってもステージ分類のようにいくつかに分かれてもよい。しかし、さらにリンパ節転移の有無（N）や遠隔転移（M）、さらにはエンドセリン染色結果を同時解析することはできない。そこでHazard ratio (HR)を用いる。これはORに時間的要素を加味したもので、複数の因子を多ロジスティック回帰で解析したように、HRを用いても複数の因子を同時に解析できる。例えば、食道癌について考えると、図10に示したように腫瘍の浸達度のレベルは予後に影響する。

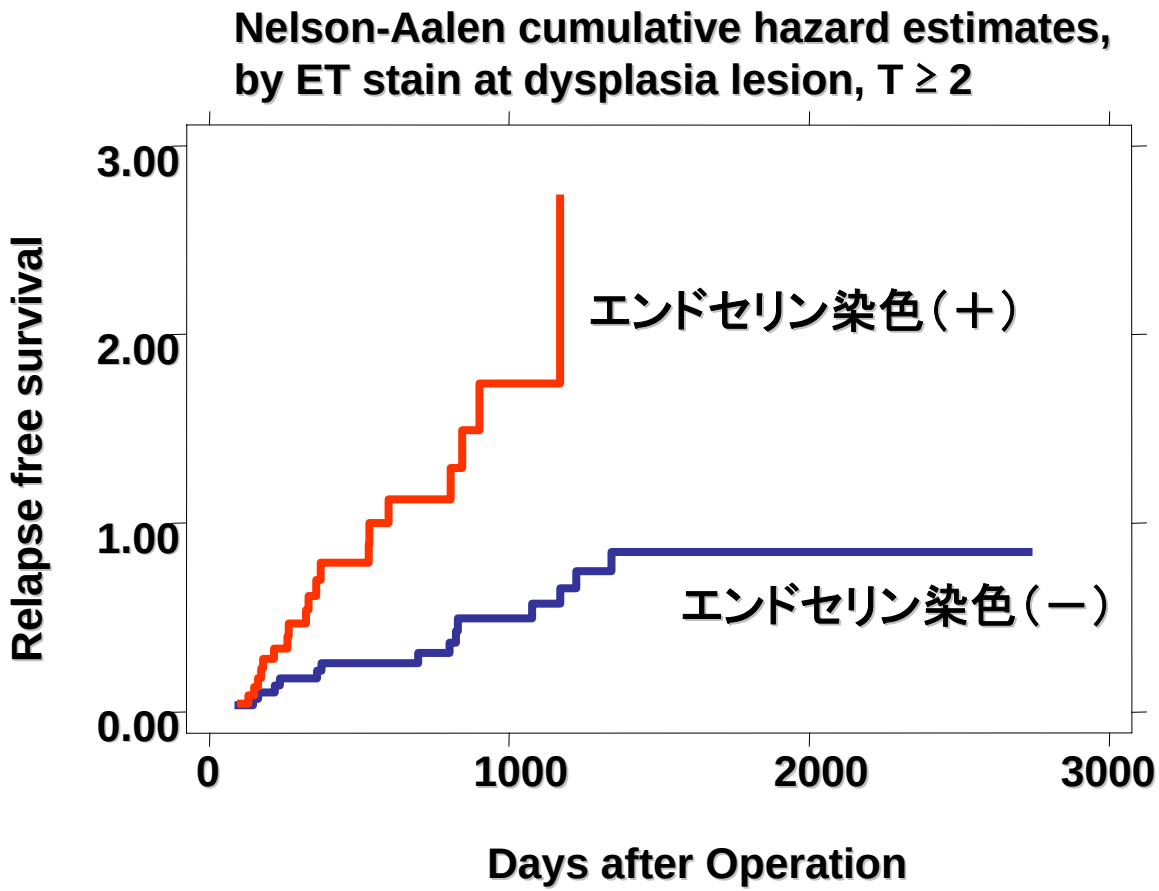


図8．食道癌患者手術症例（T1、M1を除く）のHazard 曲線
病理切片免疫染色におけるエンドセリン染色例、非染色例

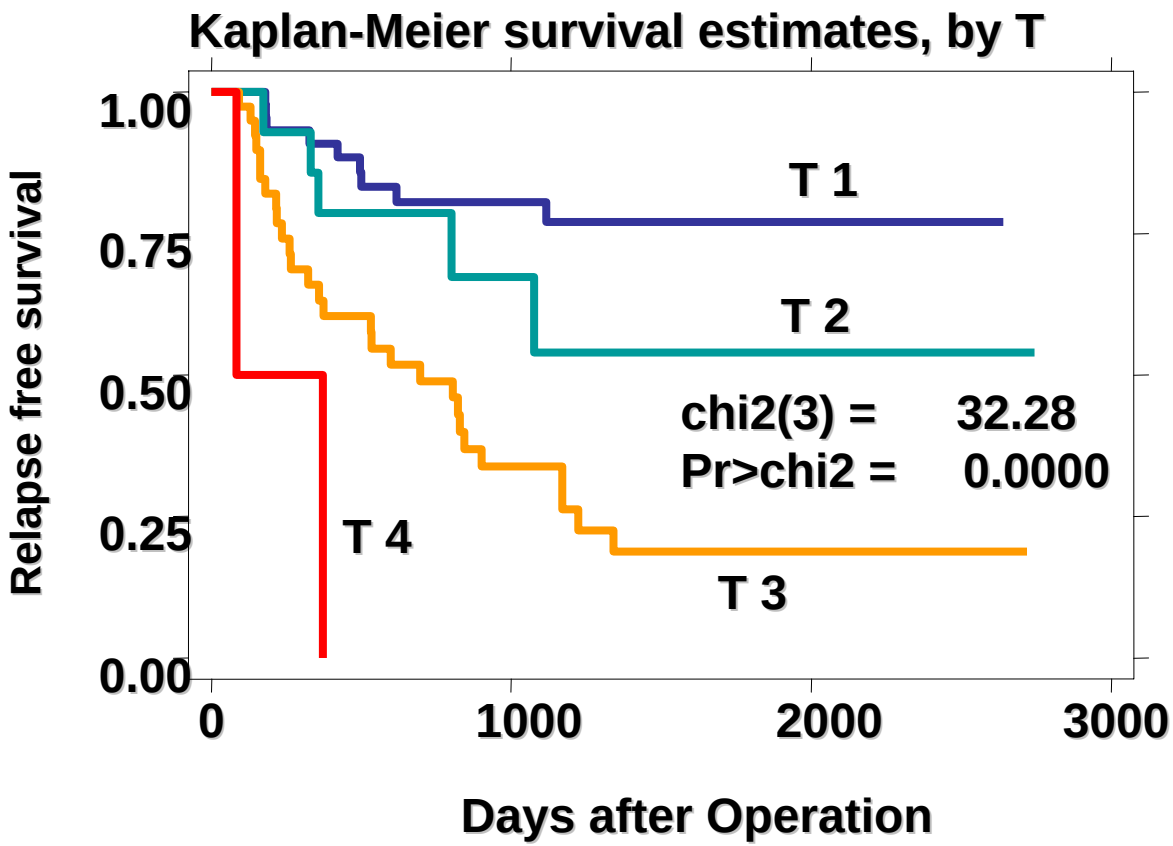


図10. 食道癌における浸達度と予後の関係

- エンドセリン染色がいくら予後に影響するとはいっても、現在知られている予後因子にエンドセリン染色の結果を加えることにより、予後予測における精度が向上しなければ、測定の意味はない。そのためには、第1にハザードモデルで既知の複数予後因子を解析する。この複数予後因子に新規予後因子を加え、新規予後因子の統計学的有意差を認め、かつモデル全体の精度の改善をみる必要がある。その1例を表6に示した。エンドセリン染色陽性のHRは2であり、95%信頼区間は1を含まないので統計学的にも有意である。さらに、エンドセリン染色を解析に加えなかったモデルと比較して、モデル全体の予測精度が改善していた(p=0.02)。よって、エンドセリン染色をTNM分類に加えて予後を判定することは意味があると結論できる(Ishibashi et al. Eur J Cancer Eur J Cancer. 2003 Jul;39(10):1409-15.)。

表６ Cox regression analysis によるエンドセリン発現と予後との関連：TNM分類で同時解析してもエンドセリンの予後との有意な関連は維持されている。

Prognostic factor	Hazard Ratio	95% CI	P value
T1 T2[*1]	**0.34**	**0.10 – 1.10**	**0.071**
T3	**1.66**	**0.62 – 4.46**	**0.318**
T4	**1.79**	**0.27 – 11.99**	**0.548**
N	**1.85**	**0.89 – 3.84**	**0.097**
M	**1.60**	**0.96 – 2.67**	**0.070**
エンドセリン染色（＋）	**2.21**	**1.10 - 4.4**	**0.026**

***1: T2 and well differentiated data dropped due to collinearity**

連続変数の際のパワー計算

今まではsuccess / failure で片が付く話でしたが、連続的な数値の場合はどうでしょうか？先と同じく世の中のAIDS患者さん全員に対してAZTを投与し 24 週後のHIV-RNAの値の平均をμ_A とし、PIを投与した場合のをμ_Bとします。よって

H_0 : $\Delta = \mu_A - \mu_B = 0$
H_A: $\Delta = \mu_A - \mu_B \neq 0$

と設定します。そして我々はAIDS全員を対象にできませんから、その極一部をとってきて全体を推論します。それぞれの治療群サンプルの平均をX_A, X_B としますと$\mu_A - \mu_B \simeq X_A - X_B$と考えられます。

$| X_A - X_B | / S\sqrt{(1/n_A + 1/n_B)} > 1.96$　　S^2 = sample variance

のときにH_0をreject します。
まだデータもないうちから sample variance が判るはずもありません。よって予想するしかありません。

$n = 2S^2 (Z\alpha + Z\beta)^2 / \delta^2$, $\delta = \mu_1 - \mu_2$

もしも両方の治療データの SD が判っていれば、

$n = (S_1 + S_2)^2 (Z\alpha + Z\beta)^2 / \delta^2$, $\delta = \mu_1 - \mu_2$

例題

脳卒中にあった患者さんで、アルファベット 24 文字が書けるまでの時間をもって回復の指標にし、2 つの治療薬を比較しようと思います。仮に両方の治療の SD が 20 秒であり、10 秒の差を検出するためにはどれくらいの脳卒中の患者さんを必要としますか？6 秒の差の場合はどうでしょうか？

解答

Δ が 10 秒のとき
$[2 \times S^2(Z\alpha + Z\beta)^2/\Delta^2] = [2 \times 20^2 \times (1.96 + 0.84)^2]/10^2 = 63$/arm

Δ が 6 秒のとき
$[2 \times S^2(Z\alpha + Z\beta)^2/\Delta^2] = [2 \times 20^2 \times (1.96 + 0.84)^2]/6^2 = 175$/arm

95％信頼区間(confidence interval)

95%信頼区間(confidence interval)は何を意味しますか？例えばあなたは研究チーフだとします。大学院生 100 人に銀座 4 丁目の交叉点を通る 300 人に年収を聞きその平均±1.96 SE を出すように指示しました。ある院生は 950 万円から 1200 万円だと述べ、ある院生は 350 万円から 900 万円だと言います。さてあなたはどの院生を信じるべきでしょうか？この 100 人の集めたデータの中に本当の値あるいは近い値が含まれているはずですが、銀座 4 丁目の交差点をその日通った人全員の本当の年収について神のみぞ知るで、あなたの知るところではありません。95 人の院生が調べたデータの範囲はまず真の平均年収をカバーするであろうと考えます。もしこの 95 人の院生の調べた範囲が 950 万円から 1000 万円だとすれ

ば、非常に正確といえますが、200 万円から 3000 万円だったとすれば、もう一度院生に年収を聞かせる方が良いかもしれません。

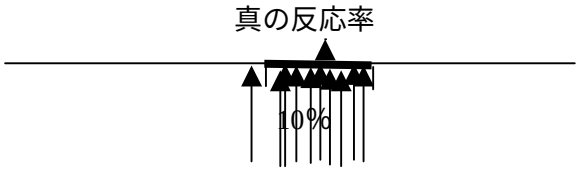
Sample size の 95%CI はどのように算出しますか？

$$(\pi_B - \pi_A) \pm 1.96 \sqrt{[\pi_A(1 - \pi_A)/n_A + \pi_B(1 \cdot \pi_B)/n_B]}$$

この部分は standard error (SE)です。

例題

phase II trial (1 arm) において、95%CI を確認したいと思います。95%の確率で、観察した反応率が真の反応率の 10%前後になるとすれば、何人の患者さんでその治療を試す必要がありますか？反応率を仮想して、それぞれ値を算出してください。



観察した反応率（100 回同じことをくり返したら 95 回は真の反応率の前後 10%に収まる）

解答

one arm なので、

$1.96 \sqrt{[\pi (1 - \pi)/n]} = 0.1$

When $\pi = 0.8$, n = 62
When $\pi = 0.7$, n = 81
When $\pi = 0.6$, n = 93
When $\pi = 0.5$, n = 96
When $\pi = 0.4$, n = 93
When $\pi = 0.3$, n = 81
When $\pi = 0.2$, n = 62
When $\pi = 0.1$, n = 35

となります。反応率が 50%のときに最も多い人数を必要とします。

Clinical Equivalence Trials

Bio-equivalence とは従来の治療薬と新しい薬を under the curve や Cmax などをもって比較するものです。これに対して Clinical Equivalence Trials とは何でしょうか？

例えばアスピリンは随分昔に開発された解熱鎮痛薬です。市場にでてから約 10 年はパテントで守られ他社が同じ薬を作って市場で売れない仕組みになっています。このパテントが切れると他社は競って類似の薬を作り出します。しかし彼らは薬の化学式のみから作るため吸収その他の面で最初に開発された薬より劣る可能性があります（もちろん優れている可能性もありますが）。そこで所謂ゾロとして発売された薬は従来の薬と効果が同じだろうかと疑問を持ちます。ゾロの薬は一般的に安いのですが、副作用さえなければ安いにこしたことはありません。このようにゾロの薬が従来の薬と比較して劣っているか同じかを比較するテストを Clinical Equivalence Trials と呼びます。よって one side で比較します。もちろんこのテストはゾロである必要はありません。作用機序が異なってもかまわないの

です。

例
例えば AZT は AIDS に対して治療効果を認められています。新たに開発された ddI は AIDS の患者さんの生存率を改善するでしょうか？

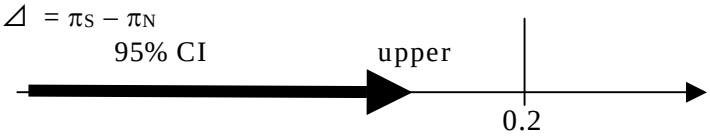
生存率の差を D とします。この差が 10%の範囲であれば同じであると考えます。

$0.1 < D < 0.1$

どうして ddI が AZT より優れた効果を考慮する必要はないのですか？今までの実験データからは AZT を超えないと予想されます。そしてこのテストの性格が同じか劣っているかを調べるのもだからです。とにかく劣ってさえいなければ OK とします。

例えば従来の治療薬(standard)が 0.7 の反応率をもち、ゾロの薬(new)が 0.5 以上であれば良しとするとします。

$\Delta = \pi_S - \pi_N$



Δ の 95%CI が 0.2 を超えていなければ OK です。

逆に 95%CI が 0.2 を超えていれば新しい薬は従来の薬より劣っていると言えます。

95%CI の上限は下記の公式で得られます。

$$(\pi_S - \pi_N) + 1.65\sqrt{[\pi_S (1 - \pi_S)/n + \pi_S (1 - \pi_S)/n]}$$

もしも真の反応率は両者で同じで、$\pi_S = \pi_N = 0.7$ であるとします。この時sample size はどれくらいになりますか？

$$1.65\sqrt{[\pi_S (1 - \pi_S)/n + \pi_S (1 - \pi_S)/n]} = 0.2$$

$$1.65\sqrt{[2 \times 0.7(1 - 0.7)/n]} = 0.2$$

$n = 29$ / arm

となります。

それではゾロの薬が 0.6 以上であれば良いとしたときどうでしょうか？
$\pi_S - \pi_N = 0.1$、

$$1.65\sqrt{[\pi_S (1 - \pi_S)/n + \pi_S (1 - \pi_S)/n]} = 0.1$$

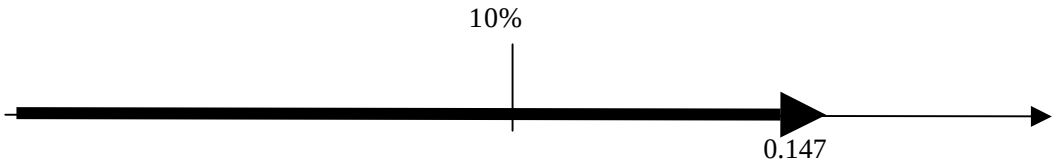
$$1.65\sqrt{[2 \times 0.7(1 - 0.7)/n]} = 0.1$$

n = 115 / arm
sample size 計算において、小さな差を検出しようと思うとより多くの人数を必要とします。
同様に差が少ないことを証明しようとすればより多くの人数を必要とします。

早期乳癌に対する simple mastectomy とより切除範囲を縮小した治療を比較したいと考えています。Simple mastectomy は既に確立した手法であり、約 80%の治癒が望めます。一方縮小腫瘍摘出術では、これで治れば患者さんにとって侵襲が少ないので好まれるのですが、治癒率が下がるのであれば本治療法を選択する理由が見当たりません。理論上縮小腫瘍摘出術は simple mastectomy を再発で超えるとは思えません。よって我々は縮小腫瘍摘出術が simple mastectomy と同じ治療成績であることを証明できればよいわけです。

我々はそれぞれ 100 人ずつの早期乳癌患者さんを縮小腫瘍摘出術と simple mastectomy とにランダムに振り分けて検討したところ、前者では 75%、後者では 80%の 5 年生存率を得ました。One sided 95%CI approach を用いて、縮小腫瘍摘出術が simple mastectomy と 5 年生存率において 10%も違わない（threshold）かどうかを検討してみてください。

$(\pi_S - \pi_N) + 1.65\sqrt{[\pi_S (1 - \pi_S)/n + \pi_S (1 - \pi_S)/n]} = (0.80 - 0.75) + 1.65\sqrt{[0.80 (1 - 0.80)/100 + 0.75 (1 - 0.75)/100]} = 0.147$



$p_1 - p_2$ の 95%CIの上限が 10%を超えてしまっているため、縮小腫瘍摘出術はsimple mastectomyと比較しnon-equivalent であると判断します。

$n_1 = [p (1 \cdot p)(1 + 1/k)(z_\alpha + z_\beta)^2]/\delta^2$
$n_2 = k \times n_1$
δ = threshold (difference)

それでは逆に両方の治療が同じ 80%の 5 年生存率を達成できると予想し、各アーム同数で検討するとし、threshold を 10%, power を 80%、a を 0.05 に設定するとすると何人について検討しなくてはなりませんか？

$n = [p (1 \cdot p)(1 + 1/k)(z_\alpha + z_\beta)^2]/\delta^2 = [0.8 \times (1 \cdot 0.8)(2)(1.645 + 0.84)^2]/0.1^2 = 198/arm$

先の例題では 100 人しか患者さんを検討しませんでした。200 人ずつで検討していたらequivalent であるという結果がでていたかもしれません。

生存曲線におけるサンプル数の計算

Hazard function の項をまず参照してください。

$\text{Prob}(T>t) = e^{-\lambda t}$

でした。ここで

H_0 : $\lambda_{1(t)} = \lambda_{2(t)}$
H_A: $\lambda_{1(t)} = \text{constant}\ \lambda_{2(t)}$

とします。2 つの治療の標準差（⊿*）をhazard rate λ_1/λ_2で示すと

$\Delta^* = \ln(\lambda_1/\lambda_2)/\sqrt{2}$

前と同じように

$d = [(Z\alpha + Z\beta)/\Delta^*]^2$　　d: は各治療における死亡数

例：疾患Xに対する現在の治療では、患者さんの生存曲線の中央値は 1 年です(λ_1)。すなわち半数は 1 年以内に死亡、半数は 1 年以上生存するということです。新しい治療で、生存曲線の中央値が 1.5 年に延びることを期待するとします(λ_2)。前と同様にα = 0.05, power = 80% と設定します。

$\Delta^* = \ln(\lambda_1/\lambda_2)/\sqrt{2} = \ln(1.5)/\sqrt{2} = 0.287$

$d = [(1.96 + 0.84) / 0.287]^2 = 96$

１つの治療アームあたり 96 人死亡があると有意差をだせそうです。病気と治療によりますが、全員が死亡するまで経過観察をしたとすると 96 人で間に合いますが、現実問題として皆死亡するわけではなくセンサーになったりサバイバーもありますからこれより多い人数が必要となります。即ち短い観察期間になればらる程、より多くの人数が必要になります。それではどれくらい必要になるのでしょうか？

		Years of additional follow-up		
		1	2	3
Years	1	150	117	104
of	2	132	110	103
accrual	3	122	107	102

Accrual とは参加者受け入れ期間のことで follow-up は参加者受け入れを打ち切ってからの観察期間を示しています。ですから accrual 1 年、follow-up 2 年といえば、合計 3 年の研究期間となります。どうやって上の数値をだしたのですか？

観察期間を考慮したサンプル数の計算

$\text{Prob}(T>t) = e^{-\lambda t}$

の公式で 1 年の平均観察期間で半数が死亡したとします（すなわち平均生存期間は 1 年、t = 1）。1 年を超えて生存する人は半数ですから、

Prob (T>1) = $e^{-\lambda 1} = 0.5$

です。これを解いて、

ln (0.5) = - 0.69

すなわち$\lambda_1 = 0.69$ となります。
元々の設定で

$\lambda_1/\lambda_2 = 1.5 = 0.69/\lambda_2 \quad \therefore \lambda_2 = 0.46$

Hazard function は小さい方が良いのです。平均生存期間が延びたことによってλ が小さくなっていますが、これで良いのです。

Accrual years = A, Follow-up years = F とします。全員が平均 F 年観察され、最初の方に登録した人と最後の方に登録した人の平均期間は A/2 です（受け入れ期間中均等に患者さんを受け入れたと仮定してです）。よって平均追跡期間は A/2 + F となります。仮に受け入れ期間（accrual）を 2 年、経過観察を 2 年としますと、平均 2/2 + 2 = 3 年となります。さてこの 3 年間は平均ですから 3 年を待たずして死亡してしまう人は全体の何%でしょうか。下記公式で

Prob (T>t) = $e^{-\lambda t}$

3 年以上生存する確率は T が failure time なので、従来の治療では、
Prob (T>3) = $e^{-\lambda t} = e^{-0.69 \times 3} = 0.126$

ですから、3 年より早期に死亡する確率は

1 - Prob (T<3) = 1 - 0.126 = 0.873

となり、一方新しい治療では、

Prob (T>3) = $e^{-\lambda t} = e^{-0.46 \times 3} = 0.252$

ですから、3 年より早期に死亡する確率は

1 - Prob (T<3) = 1 - 0.252 = 0.748

となります。

さて上で１アーム当り 96 人の死亡が必要であると計算されました。統計学者は臨床試験を解析するにあたって sample size よりもより多くの event を期待するのです。さて、すぐ上の計算式から、従来の治療では 3 年満期を待たずして 87.3%の人が死亡することが予想されます。一方新しい治療では 74.8%です。新しい治療の方が有効であろうと予測していますから、納得いく数値です。統計学者は 96 人の死亡が必要だといっていますから、96 人がそれぞれのアームで 87.3%, 74.8% に相当すれば良いわけですから、最初に必要な人数(sample size)は 110 人と 128 人であり、合計 238 人となります。上の表に近い値となりました。年間何人位参加者を募るか予想ができれば、本当にその accrual でよいかどうか検討できます。上のような表を作って計画をたてるとやりやすいかもしれません。

例題
AIDS の患者さんの従来の治療における平均生存期間は 1.5 年だとします。もしも新しい治療では 2 年間の平均生存期間を期待できるとします。さてこの２つの治療において randomized clinical trial を行なう予定にしていますが、何人の AIDS 患者さんの参加をつのればよいでしょうか？参加する AIDS 患者さんが年間 300 人（各アーム 150 人）であり、3 年間受け入れ期間(accrual)を設定し、1 年間経過観察するとします。Type I error 5%, type II error 20% として計算してみてください。

解答

$\text{Prob}\,(T>t) = e^{-\lambda t}$

でした。ここで

H_0 : $\lambda_{1(t)} = \lambda_{2(t)}$
H_A: $\lambda_{1(t)} = \text{constant}\ \lambda_{2(t)}$

とします。

$\Delta^* = \ln(\lambda_1/\lambda_2)/\sqrt{2} = \ln(2.0/1.5)/\sqrt{2} = 0.203$

$d = [(1.96 + 0.84) / 0.203]^2 = 190.24$

１つの治療アームあたり 190 人死亡があると有意差をだせそうです。病気と治療によりますが、全員が死亡するまで経過観察をしたとすると 190 人で間に合いますが、現実問題として皆死亡するわけではなくセンサーになったりサバイバーもありますからこれより多い人数が必要となります。即ち短い観察期間になればらる程、より多くの人数が必要になります。それではどれくらい必要になるのでしょうか？

$\text{Prob}\,(T>t) = e^{-\lambda t}$

の公式で 1.5 年の平均観察期間で半数が死亡しますから、

$\text{Prob}\,(T>1.5) = e^{-\lambda \times 1.5} = 0.5$

です。これを解いて、

$\ln(0.5) = -\lambda \times 1.5\ ,\ \lambda = -0.46$

すなわち$\lambda_1 = 0.46$ となります。
元々の設定で

$\lambda_1/\lambda_2 = 2.0/1.5 = 1.33 = 0.46/\lambda_2 \quad \therefore \lambda_2 = 0.35$

受け入れ期間（accrual）を 3 年、経過観察を 1 年としますと、平均 3/2 + 1 = 2.5 年となります。さてこの 2.5 年間は平均ですから 2.5 年を待たずして死亡してしまう人は全体のどれくらいにあたるでしょうか。下記公式で

$\text{Prob}\,(T>t) = e^{-\lambda t}$

2.5 年以上生存する確率は T が failure time なので、従来の治療では、
Prob (T>2.5) = $e^{-\lambda t} = e^{-0.46 \times 2.5} = 0.317$

ですから、2.5 年より早期に死亡する確率は

1 - Prob (T<2.5) = 1 - 0.317 = 0.683

となり、一方新しい治療では、

Prob (T>2.5) = $e^{-\lambda t} = e^{-0.35 \times 2.5} = 0.417$

ですから、2.5 年より早期に死亡する確率は

1 - Prob (T<2.5) = 1 - 0.417 = 0.583

となります。

さて上で１アーム当り 190 人の死亡が必要であると計算されました。統計学者は臨床試験を解析するにあたって sample size よりもより多くの event を期待するのです。さて、すぐ上の計算式から、従来の治療では 3 年満期を待たずして 68.3%の人が死亡することが予想されます。一方新しい治療では 58.3%です。新しい治療の方が有効であろうと予測していますから、納得いく数値です。統計学者は 190 人の死亡が必要だといっていますから、190 人がそれぞれのアームで 68.3%, 58.3% に相当すれば良いわけですから、最初に必要な人数 (sample size)は 278 人と 326 人であり、合計 604 人となります。

STATA を用いたsample size の計算

例題１.
狭心症の新薬について randomized placebo controlled clinical trial を行なうことになりました。薬効評価については、randomization を行なった時点と、治療薬を開始して 4，6，8 週後に運動負荷試験を行なって胸痛が出現するまでの時間（秒）で測定しようと思います。以前に行なった pilot study では placebo 群で 498 ± 20.2 sec, 薬剤投与群で 485 ± 19.5 sec でした。経過観察中の相関を 0.7 とします。個々の患者さんの治療開始前後での変化をみるのでこれを change method と呼ぶことにしましょう。$\alpha = 0.05$ (two sided), 90% power で条件設定をしたとき、何人の患者さんが必要でしょうか？

STATA の command に以下のようにタイプしてみてください。

```
. sampsi 498 485, sd1(20.2) sd2(19.5) method(change) pre(1) post(3)
r1(.7)

Estimated sample size for two samples with repeated measures
Assumptions:
        alpha =   0.0500  (two-sided)
                                power =   0.9000
                                   m1 =      498
                                   m2 =      485
                                  sd1 =     20.2
                                  sd2 =     19.5
                                n2/n1 =     1.00
          number of follow-up measurements =       3
 correlation between follow-up measurements =   0.700
           number of baseline measurements =       1
   correlation between baseline & follow-up =   0.700

Method: CHANGE
 relative efficiency =    2.500
    adjustment to sd =    0.632
        adjusted sd1 =   12.776
        adjusted sd2 =   12.333

 Estimated required sample sizes:
                n1 =      20
                n2 =      20
```

薬剤投与群、placebo 群それぞれ 20 人となりました。上のような繰り返し測定する場合には複雑な計算が必要であり、コンピュータを用いた計算がとても便利です。

Clinical trials with repeated measures　（治療前後での比較）

我々は 30 人を検討する分のグラントしかないとします。それでも統計学的に検討できるでしょうか？１アームの人数は 15 人になります。

```
. sampsi 498 485, sd1(20.2) sd2(19.5) method(change) pre(1) post(3)
r1(.7) n1(1
5) n2(15)

Estimated power for two samples with repeated measures
Assumptions:
                                alpha =   0.0500  (two-sided)
                                   m1 =     498
                                   m2 =     485
                                  sd1 =    20.2
                                  sd2 =    19.5
                           sample size n1 =      15
                                   n2 =      15
                                 n2/n1 =    1.00
         number of follow-up measurements =       3
 correlation between follow-up measurements =    0.700
          number of baseline measurements =       1
   correlation between baseline & follow-up =    0.700

Method: CHANGE
 relative efficiency =    2.500
    adjustment to sd =    0.632
       adjusted sd1 =   12.776
       adjusted sd2 =   12.333

 Estimated power:
             power =    0.809
```

.

80%のパワーがあります。まずまずの数値です。それでは 30 人で検討することにしましょう。この薬剤は placebo より効果が期待できるかもしれません（定かではないから試験をするわけですが、薬剤使用アームを増やした方が患者さんをリクルートしやすい利点があります）。薬剤投与群を 20 人にしたらどうでしょうか？

```
. sampsi 498 485, sd1(20.2) sd2(19.5) method(change) pre(1) post(3)
r1(.7) n1(20) n2(15)

Estimated power for two samples with repeated measures
Assumptions:
                                alpha =   0.0500  (two-sided)
                                   m1 =     498
                                   m2 =     485
                                  sd1 =    20.2
                                  sd2 =    19.5
                           sample size n1 =      20
                                   n2 =      15
                                 n2/n1 =    0.75
         number of follow-up measurements =       3
 correlation between follow-up measurements =    0.700
```

```
          number of baseline measurements =       1
   correlation between baseline & follow-up =   0.700

Method: CHANGE
 relative efficiency =    2.500
    adjustment to sd =    0.632
        adjusted sd1 =   12.776
        adjusted sd2 =   12.333

 Estimated power:
             power =    0.860
```

.

86%のパワーがあります。

Two-sample test of equality of proportions　（Yes/no type の試験）

インフルエンザ罹患率は 10%とします。新しい予防薬が開発されこれを内服することにより 3%まで減少させることが期待されるとします。この 10%と 3%が違うか同じかは sample size によるわけですが、新薬の効果がないとするH_0 をreject するには$\alpha = 0.05$, power 0.80 とした場合どれくらいのsample 数が必要でしょうか？

```
. sampsi 0.1 0.03, power(0.8)

Estimated sample size for two-sample comparison of proportions

Test Ho: p1 = p2, where p1 is the proportion in population 1
                and p2 is the proportion in population 2
Assumptions:

        alpha =   0.0500  (two-sided)
        power =   0.8000
           p1 =   0.1000
           p2 =   0.0300
        n2/n1 =   1.00

Estimated required sample sizes:

           n1 =     222
           n2 =     222
```

1 アーム 222 人必要です。この薬剤は phase I trial にて比較的安全な薬であることがわかっています。パワーを 90%まで上げるとどうなりますか？

```
. sampsi 0.1 0.03, power(0.9)

Estimated sample size for two-sample comparison of proportions

Test Ho: p1 = p2, where p1 is the proportion in population 1
                and p2 is the proportion in population 2
Assumptions:

        alpha =   0.0500  (two-sided)
        power =   0.9000
           p1 =   0.1000
           p2 =   0.0300
        n2/n1 =   1.00

Estimated required sample sizes:

           n1 =     287
           n2 =     287
```

１アーム 287 人必要になります。

さて新薬の方を多く設定したいと思います。例えば薬剤投与を 300 人、placebo を 150 人に設定したとすると、

```
. sampsi 0.1 0.03, n1(300) r(0.5)

Estimated power for two-sample comparison of proportions

Test Ho: p1 = p2, where p1 is the proportion in population 1
                  and p2 is the proportion in population 2
Assumptions:

        alpha =   0.0500  (two-sided)
           p1 =   0.1000
           p2 =   0.0300
sample size n1 =     300
           n2 =     150
        n2/n1 =    0.50

Estimated power:

        power =   0.7185
```

.
患者さんの総数はあまり変わらなくてもパワーが落ちてしまいます。同じ人数の時、パワーはそれぞれのアームの人数が同じ時最も強くなります。

それでは薬剤と placebo の関係を 2:1 に保ったまま 80%のパワーで検討するためには何人が必要となりますか？

```
. sampsi 0.1 0.03, power(0.8)  r(0.5)

Estimated sample size for two-sample comparison of proportions

Test Ho: p1 = p2, where p1 is the proportion in population 1
                  and p2 is the proportion in population 2
Assumptions:

        alpha =   0.0500  (two-sided)
        power =   0.8000
           p1 =   0.1000
           p2 =   0.0300
        n2/n1 =   0.50

Estimated required sample sizes:

           n1 =     349
           n2 =     175
```

One sample test of proportion　（従来の治療と比較する）

ある疾患に対してステロイドパルス療法を行なったところ 8 人治療して 6 人が寛解に入りました。さて従来の治療とこれからの治療を比較するとしましょう。その疾患に対するステロイドの寛解率は 50%であり、どの教科書をみても同じ数値なので golden standard として用いることができるとします。さて我々はパルス療法の効果が通常のステロイド療法より効果があるかどうか調べたいのですが、仮に 75%の寛解率を得るとして、 a = 0.05, 80%のパワーをもって証明するためには何人の患者さんにパルス療法を施行しなくてはなりませんか？

```
. sampsi 0.5 0.75, power(0.8) onesample

Estimated sample size for one-sample comparison of proportion
  to hypothesized value

Test Ho: p = 0.5000, where p is the proportion in the population

Assumptions:

        alpha =   0.0500  (two-sided)
        power =   0.8000
 alternative p =   0.7500

Estimated required sample size:

            n =      29
```

29 人です。しかしこのような比較は historical comparison と呼ばれ randomized clinical trial と比較すると信頼性が低くなります。特に新しい治療と従来の治療の差が小さい時はいくら有意差があるといっても周りを説得することはできません。

```
. sampsi 0.5 0.75, power(0.8) onesample

Estimated sample size for one-sample comparison of proportion
  to hypothesized value

Test Ho: p = 0.5000, where p is the proportion in the population

Assumptions:

        alpha =   0.0500  (two-sided)
        power =   0.8000
 alternative p =   0.7500

Estimated required sample size:

            n =      29
```

Two sample test of equality of means　（連続変数をendpoint とした試験）

我々は抗高血圧薬の効果を調べようと思います。その対象となる患者さんの平均拡張期血圧は 105 mmHg であり、SD は 10 mmHg だとします。そしてこの薬剤により 98 まで下がると想定します。SD に関しては全くデータがないため母集団と同じ 10 とします。薬剤使用群と placebo 群の比を 2:1 で比較するにはそれぞれのアームで何人が必要となりますか？パワーは 80%、$\alpha = 0.05$ とします。

```
. sampsi 105 98, p(0.8) r(2) sd1(10) sd2(10)

Estimated sample size for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                 and m2 is the mean in population 2
Assumptions:

        alpha =   0.0500  (two-sided)
        power =   0.8000
           m1 =      105
           m2 =       98
          sd1 =       10
          sd2 =       10
        n2/n1 =     2.00

Estimated required sample sizes:

           n1 =       25
           n2 =       50
```