Sample size と power calculation

Sample Size Estimation

例えば、AZTとPIのAIDSに対する治療効果を検討しようと思います。何人になるか判りませんが、世の中全員のAIDSの患者さんに対してAZTを投与して効果をみます。また同じく全員のAIDS患者さんに対してPIを投与して効果をみます。ここで、それぞれの患者さんは服薬によってAIDSウイルスRNAが検出限界以下になったら治療成功とします。ここでAZTの成功する確率Pr (S_A) = π_A, PIの成功する確率Pr (S_B) = π_Bとします。しかし現実問題世の中のAIDSの患者さん全部を集めて治療することは不可能です。そこで、一部の患者さんのデータ(sampling)から全体を推論(inference)することになります。π_A , π_Bは推論したものであり、文字の上にハットをつけるルールですが、本書では省略します。

もしもπ_A - π_B = 0.20 だったとします。PIの方が 20%も治るとすれば、画期的です。しかし、この 20%は本当に違うのでしょうか？AZT, PIそれぞれ 10 人ずつ治療して、6 人と 8 人HIV-RNAが検出されなくなっただけだとします。これではいくら 20%の違いといっても差があるとはいえません。それではどのような条件のとき差があるといえるのでしょうか？

n_A = AZT で治療を受ける患者さんの数、S_A = AZTで治療を受けて成功する数、n_B = PIで治療を受ける患者さんの数、S_B = PIで治療を受けて成功する数、とすれば、

$\pi_A = S_A/n_A$,　$\pi_B = S_B/n_B$, となります。知りたいのは世界中のAIDS患者さんに対する治療効果の差（= Δ）ですが、これはδ　=　π_A　-　π_B　で代用されます。

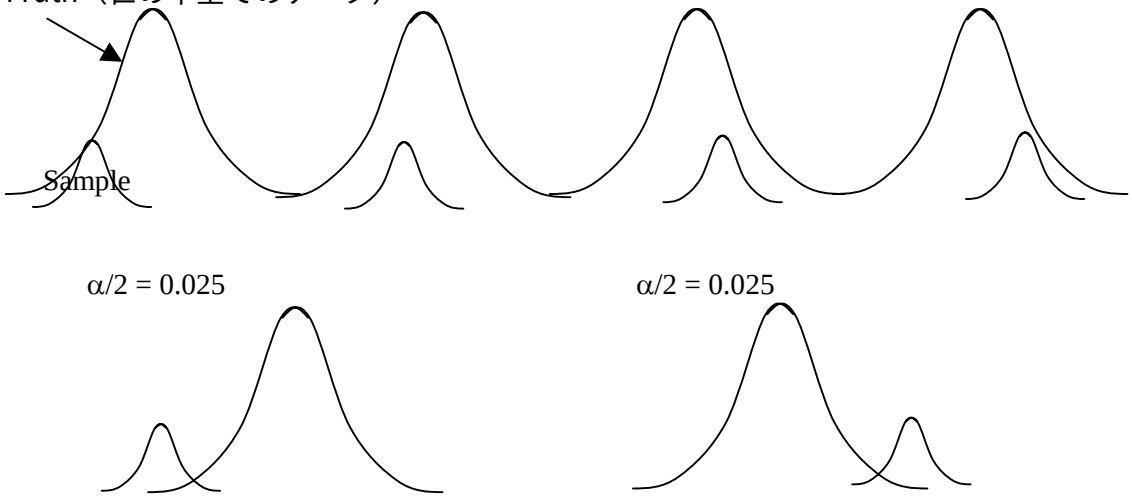
$H_0 : \Delta = 0$
$H_A : \Delta \neq 0$

もしも$|\delta|$が非常に大きかったらH_0 をreject してH_Aをaccept できます。一体どれくらい大きければよいのですか？

$|\delta|/SE > 1.96$ の時に reject できます。

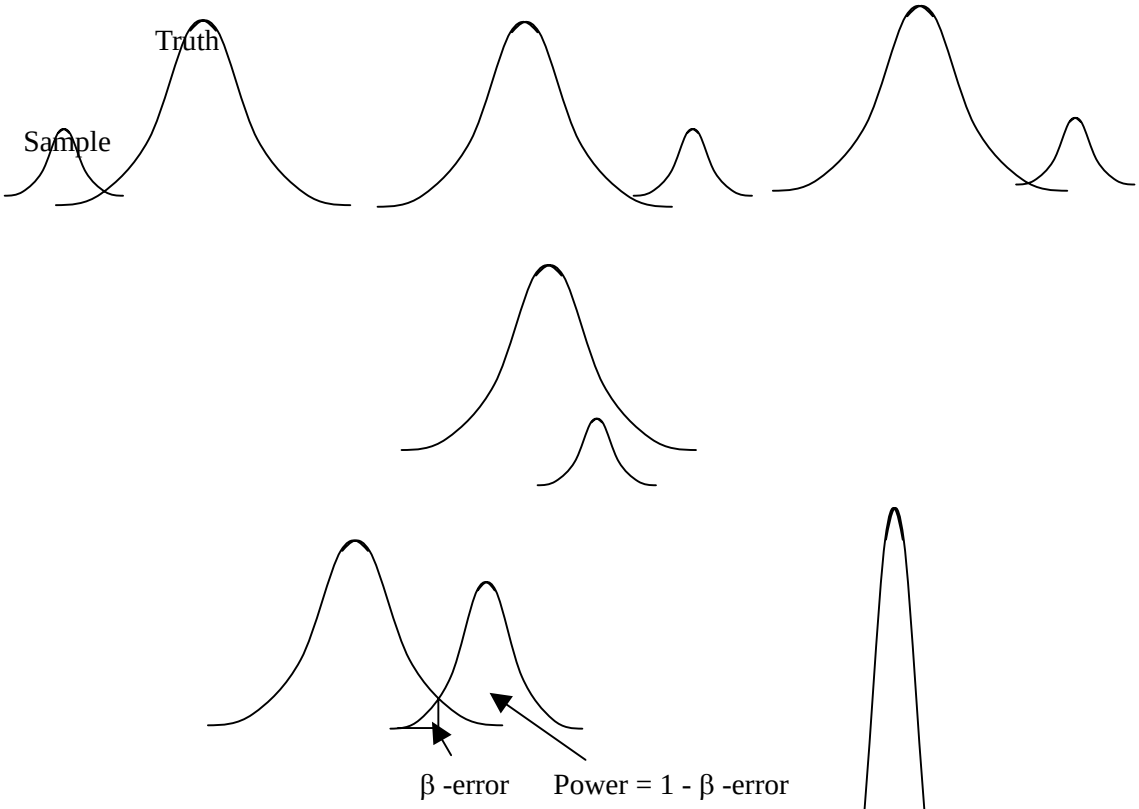
$SE = \sqrt{\{\pi_A(1-\pi_A)/n_A\} + \{\pi_B(1-\pi_B)/n_B\}}$

$|\delta|/SE$が 2.0 でH_0をrejectしました。これはどういうことですか？5% の確率で間違うこともある、すなわち本当はH_0が正しいのに（差がないのに）差があると言ってしまう（間違ってしまう）確率が 5%はあるということです。20 回の同じ臨床試験を行なうと 1 回は間違うこともあるとも言えます。

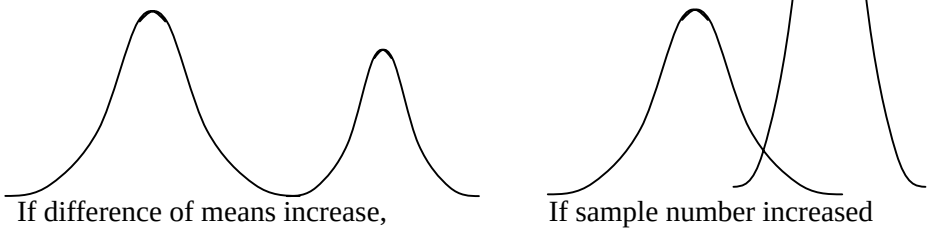


			God	
			Truth	
			H_0	H_A
You	Sample	H_0	A=B (OK)	Type II error (β error)
		H_A	Type I error (α error)	A is not B

神様とあなたは違います。何故なら神様は常に truth を述べますが、あなたは人間ですから間違うこともあります。本当は A と B が同じなのに、あなたが間違って A と B は違うといってしまうことを Type I error(α error)といい、逆に本当は A と B が違うのにあなたは間違って同じだといっていますことを Type II error(β error)といいます。FDA などは効かない薬を効くといってもらっては困るのでα error を気にします。一方製薬会社は効かない薬を効くといって売っても商売になるので気にしません。しかし、本当は効くのに効かないと間違われては、今までかけてきた時間とお金が無駄になってしまうため、β error を慎重に検討します。



power はどのようなときに大きくなりますか?



β-error は図からみてわかる通り one side です。何故なら 2 つは異なると仮定しており、２つの交わりは１つでしかあり得ないのです。

それでは最初の AIDS 治療の話題に戻りましょう。仮に AZT で 20%が、PI で 30%が HIV-RNA 陰性化するとします。そしてα = 0.05 (two sided), β = 0.20 (power = 80%)と設定します。これは臨床試験においてしばしばみられる設定パターンです。α = 0.05 (two sided)はほぼ固定しています。

$\delta = (0.3 - 0.2)/\sqrt{0.3\,(1 - 0.3) + 0.2\,(1 - 0.2)} = 0.1644$

$n = [(Z\alpha + Z\beta)/d]^2 = [(1.96 + 0.84)/0.1644]^2 = 290$

つまり AZT 治療群 290 人と PI 治療群 290 人に分配することを考え、合計 580 人の患者さんをリクルートします。

例題

あなたは２つの治療を比較しようと思っています。Power を 90%、two sided test α = 0.05, として、反応率が 35% から 45%に改善する場合、50%から 60%に改善する場合にそれぞれ何人の患者さんを募らなくてはならないでしょうか？同じ 10%の改善を検出するわけですが、標本数は同じですか？

解答

35% ↦ 45%

(0.45 – 0.35) / $\sqrt{}$(0.45 x 0.55 + 0.35 x 0.65) = 0.145

[(1.96 + 1.28) / 0.145] = 500 人/アーム

50% ↦ 60%

(0.60 – 0.50) / $\sqrt{}$(0.60 x 0.50 + 0.50 x 0.50) = 0.143

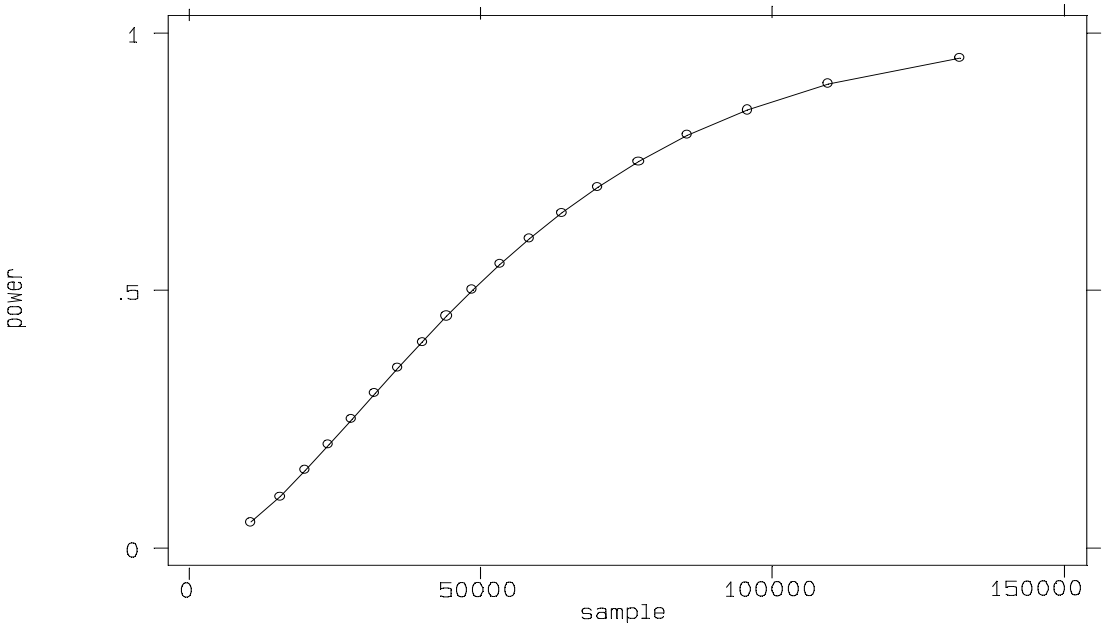
[(1.96 + 1.28) / 0.143] = 513 人/アーム

ポリオワクチン臨床試験

1950 年代前半ポリオウイルスに関する知見が集まり、ワクチンが開発されました。そして National Foundation for Infantile Paralysis (NFIP) 主導のものと、かつ歴史的といってもいい程大掛かりな臨床試験が行なわれることになりました。最初は小学校 2 年生に対してワクチンを施行し、1 年と 3 年生には何もしないで経過を観察する計画をたてました。一方 double blind placebo-controlled study も可能であり、むしろその方が良いのではないかとの意見もあり、ミシガン大学ワクチン評価センターの Thomas Fancis Jr. 博士のもと両方の臨床試験が推し進められることになりました。

ポリオはその年、場所によって流行状況が大きく異なるため、比較的頻度の多い場所を適当に選んで行なわれることになりました。およそ 10 万人あたり 30 人のポリオによる麻痺が発生するとして、どのくらいの子供を対象にする必要があるでしょうか？α = 0.05 (two sided), power = 5%から 95%, ワクチンにより 90%改善する、すなわち発生数が 0.10 になると仮定して sample size を計算してみてください。

(www.mc.vanderbilt.edu/prevmed/ps.htmからsample size 計算用ソフトを読み取ることができます。)。ワクチンが不活化されていなければ副作用としてポリオ麻痺を一般よりも多く発生させてしまう可能性も残っています。ですからone-sided ではなく two-sided にするべきです。また本当のワクチンを投与する群とプラシーボ群の関係は 1：1 とします。Power を変化させたときのsample size の変化を下に示します。



この図からわかる通りパワーと sample size の関係は直線にはなりません。パワー5%のときの sample size を倍にするとパワーは 18%まで上昇します。パワー50%のときの sample size を倍にするとパワーは 35%まで上昇します。パワー90%のときの sample size を倍にするとパワーは 99%まで上昇します。つまりパワーは 50%前後のとき sample size に強く影響を受けると言えます。結局 10 万人から 15 万人を対象にすると結果を推論する際、十分なパワーが得られそうです。

ワクチンの安全性

ワクチン開始前、フォルマリン処理で大丈夫か、株の選択はこれでよいか、今までの安全性試験の信頼性などについて異論がありました。特に物議を醸し出した異論は「ワクチンは完全に不活化できず生きたウイルスを注射することになる可能性があるため危険である」というものでした。これに対してロックフェラー大学 Thomas M Rivers 博士らを中心とする委員会が組織され、全てのワクチン製造過程のマニュアルを検討し、更に３つの研究施設にワクチン不活化が不完全でないかどうか調べさせました。このように入念な安全性の検討の末 1954 年 4 月臨床試験が開始されました。初回投与、1 週間あけて第二回目投与、そして 2 回目から 4 週間後に３回目の注射を行ないます。44 週、211 の地域、以下の人数の子供が参加しました。このような大掛かりな臨床試験は現代でもまずありません。

Double blind placebo-controlled trial

参加状況	小学校１，２、３年	
	実数	%
合計	749,236	100.0
参加	455,474	60.8
完全ワクチン注射	200,745	26.8
完全プラシーボ注射	201,229	26.9
不完全ワクチン注射	8,484	1.1
不完全プラシーボ注射	8,577	1.1
欠席	36,439	4.9
参加同意を得られず	280,868	37.5
参加の記載がない	12,894	1.7

学年毎の研究

参加状況	小学校１，２、３年		2 年		1 年、3 年	
	実数	%	実数	%	実数	%
合計	1,080,680	100	355.507	100	725,173	100
参加	567,210	52.5	245,895	69.2	321,315	44.3
完全ワクチン注射	221,998	20.5	221,998	62.4		
不完全ワクチン注射	9,904	0.9	9,904	2.8		
欠席	13,993	1.3	13,993	3.9		
参加同意を得られず	332,870	30.8	105,211	29.6	227,659	31.4
参加の記載がない	180,600	16.7	4,401	1.2	176,199	24.3

下の表で 1 年と 3 年はワクチンを接種せずに観察するだけですから、どうしても観察者側の目が行き届きません。2 年生の記録保存はしっかりとしていますが、1 年と 3 年は 3 人に 1 人で失われています。そういう人に限ってポリオになりやすかったり、なりにくかったりするとデータのねじれを生じバイアスのもとになります。そこで参加者と非参加者の間に何か関連がないかどうか検討してみたところ、非参加者は家族の収入、教育などが低い傾向にありました。このことは、非参加者の方がよりポリオウイルスに暴露されやすい、すなわちワクチン試験開始時ポリオに免疫を持つ人が多かった可能性を示唆しています。実際非参加者のデータも含めて考えると、RR は 0.5 となり予防硬

化が落ちています。注射後発疹、喘息なども僅かながら観察されましたが、プラシーボと変わらなかったため、明らかなワクチンの副作用は認められませんでした。また学校の保健室にはどんな些細な症状でも書きとめて欲しいと依頼してあり、その結果は下の表です。

	受けた数	小さな症状		やや大きな症状	
		実数	%	実数	%
blind					
ワクチン	209,229	931	0.4	9	0.004
プラシーボ	209,806	939	0.4	13	0.006
小学校 2 年					
ワクチン	231,902	1,694	0.7	7	0.003

小さな症状もやや大きな症状もワクチン投与群とプラシーボ群で大きな変わりはありません。よってこれらの症状はポリオワクチンとは直接関係ないと思われます。小さな症状は自分あるいは親がワクチンを接種されているとわかっている群で多い傾向にありました。これは特に小さな副作用でしばしば認められる現象であり、だからこそプラシーボを必要とするのです。

最大の問題はワクチンの不活化不十分によりポリオ麻痺を来たすことはないかどうかです。そこで第１回目ワクチン接種から第３回目ワクチン接種までの間の麻痺発生を検討しています。

Double blind placebo-controlled trial

	投与数	麻痺発生数	10 万人当りの発生数
ワクチン群	209,229	4	1.9
プラシーボ群	209,806	5	2.4
他	330,201	10	3.0

ワクチンとプラシーボ群で麻痺発生に差を認めません。この表をみるかぎりワクチン不活化不十分の問題はクリアしています。学年毎で比較した研究ではどうでしょうか？

	投与数	麻痺発生数	10 万人当りの発生数
ワクチン群（2 年）	231,902	11	4.7
コントロール群（1 年および 3 年）	725,173	37	5.1
他	123,605	4	3.2

やや学年毎に施行した表において麻痺の頻度が多いようですが、実際にはワクチン接種の有無と麻痺との間には相関関係を見出せませんでした。

結局のところ全部で 1,012 人のポリオ患者を認め、そのうち 428 人が double blind placebo-controlled trial において発生し、584 人が学年でワクチン投与群を観察群で分けた試験において発生しました。内訳は以下の如くです。

Placebo-controlled trial

	患者数			10 万人当りの患者数			
	ワクチン	プラシーボ	他	ワクチン	プラシーボ	他	RR

麻痺	33	110	124	16	55	36	0.29
非麻痺	23	28	37	11	14	11	0.79
ポリオ疑	10	7	7	5	3	2	1.67
非ポリオ	15	17	17	7	8	5	0.88
合計	81	162	185	40	81	53	0.49

非麻痺に関しては差ほどではありませんが、麻痺を起こす患者数はプラシーボ群と比較してワクチン投与群で 1/3 にまで抑えられています。

Observed control

	患者数			10 万人当りの患者数			
	ワクチン	コントロール	他	ワクチン	コントロール	他	RR
麻痺	38	331	46	17	46	34	0..37
非麻痺	17	60	11	8	8	8	1.00
ポリオ疑	12	24	6	5	3	4	1.67
非ポリオ	8	25	6	4	3	4	1.33
合計	75	440	69	34	61	52	0.56

この試験においてもワクチンはポリオ麻痺発生を抑制しています。年齢による effect modification も重要な所見でした。6 歳では 24%の抑制（有意差なし）でしたが、7 才では 75%、8 歳では 87%、9 歳では 89%のポリオ麻痺に対する予防効果がありました。

これがもしも信頼性の高い臨床試験でなかったら、本当のところワクチンの効き目について知ることはできなかったでしょう。この歴史的ポリオ・ワクチン臨床試験は、その後の臨床試験のあり方を大きく変えたことは言うまでもありません。

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

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

⊿ 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。

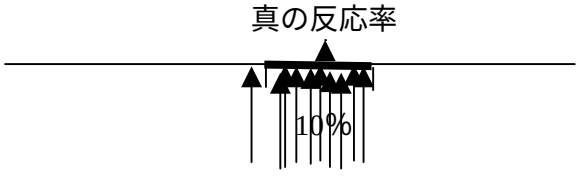
Sample size の 95%CI はどのように算出しますか？

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

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？

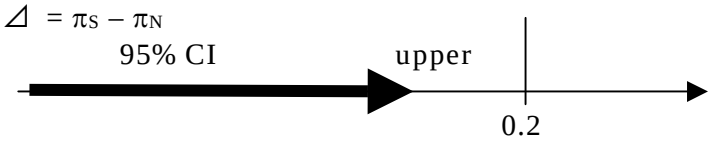
生存率の差を D とします。この差が 10%の範囲であれば同じであると考えます。

$0.1 < D < 0.1$

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。



⊿ 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？
$\pi_S - \pi_N = 0.1$、

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

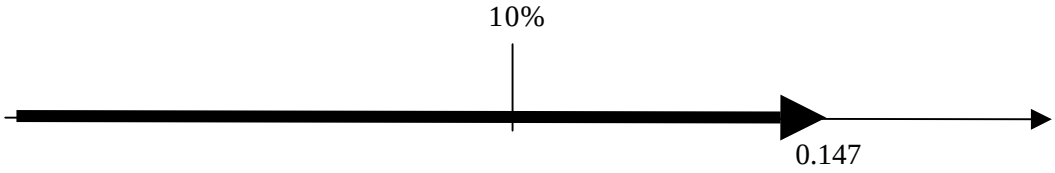
$$1.65\sqrt{[2 \times 0.7(1-0.7)/n]} = 0.1$$

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。

$n_1 = [p (1 - p)(1 + 1/k)(z_\alpha + z_\beta)^2]/\delta^2$
$n_2 = k \times n_1$
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？

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

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。
Prob (T>t) = $e^{-\lambda t}$

でした。ここで

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

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、

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

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

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

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、
$\text{Prob}\ (T>3) = e^{-\lambda t} = e^{-\ 0.69 \times 3} = 0.126$

ですから、3 年より早期に死亡する確率は

$1 - \text{Prob}\ (T<3) = 1 - 0.126 = 0.873$

となり、一方新しい治療では、

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

ですから、3 年より早期に死亡する確率は

$1 - \text{Prob}\ (T<3) = 1 - 0.252 = 0.748$

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、

$\ln(0.5) = -0.46$

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、

$$\text{Prob}\,(T>2.5) = e^{-\lambda t} = e^{-0.46 \times 2.5} = 0.317$$

ですから、2.5 年より早期に死亡する確率は

$$1 - \text{Prob}\,(T<2.5) = 1 - 0.317 = 0.683$$

となり、一方新しい治療では、

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

ですから、2.5 年より早期に死亡する確率は

$$1 - \text{Prob}\,(T<2.5) = 1 - 0.417 = 0.583$$

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 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、96 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。α = 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α = 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
```