例数設計の基礎

第 8 回 Armitage 勉強会

土居正明

1　はじめに

1.1　本稿の内容

本稿では、2 群間の平均を比較する t 検定の例数設計についてご説明します[*1]。

例数設計を理解する際、最も大事なことは α エラーと β エラーの 2 つのエラーをきちんと理解しておくことです。ですので、まず最初にこれらエラーから見ていきましょう。

1.2　用語の確認

一つ、大きな混乱のもととなる用語を整理しておきます。それは**「平均」**という言葉です。本稿で**「平均（値）」**と言えば母集団の平均 μ（**未知**の値）を指すものとします。そして、**「標本平均」**というと、データを足して例数で割った統計量 $\bar{x}$（つまり μ の推定値・**既知**の値）を指すものとします。どちらの話をしているのかを間違えてしまいますと大変混乱しますので、よく注意してください。

また、今回**「分布」**には データの分布 と 標本平均の分布 の 2 種類が出てきます。この 2 つをしっかり区別しながら読んでください。

2　α エラーと β エラーの話

2.1　言葉の準備

まず最初に、以下の表をしっかり理解しておきましょう。

表 1　α エラーと β エラーの定義

		検定結果	
		「差がない（効かない）」と判断	「差がある（効く）」と判断
現実	差がない (効かない薬)	正しい判断	α エラー （企業の不当な利益）
	差がある （効く薬）	β エラー （企業の不当な損失）	正しい判断

[*1] 厳密には「近似的な方法」ですが、かなり精度のよい近似になっています。t 検定では分散は推定値を使いますが、今回は既知の値を用いる点が異なります。結果として、検定に t 分布は必要なく、正規分布で十分となります。

α **エラー**：「効かない」薬を「効く」と判断するので、企業に有利な間違い です。
従って、**当局はこちらを小さくするよう要請します**。
β **エラー**：「効く」薬を「効かない」と判断するので、企業に不利な間違い です。
従って、**企業はこちらを小さくしたいと思います**[*2]。

一般に、α **エラーが起こる確率を** α**、**β **エラーが起こる確率を** β **で表します。**

さて、これを受けてさらに2つの言葉を導入しましょう。
有意水準：「効かない」薬があったときに「効く」と判断してしまう確率。**（**α **エラーを起こす確率）**$= \alpha$。
検出力：「効く」薬があったときに「効く」と判断できる確率。$1 -$（β **エラーを起こす確率）**$= 1 - \beta$。
上の表と見比べると、有意水準は小さい方がよく、検出力は大きい方がよい、ということになります。

2.2 あっちが立てばこっちが立たず

では、「有意水準（α エラーを起こす確率）を 0 にして、検出力（β エラーを起こさない確率）を 1 にしたい」と思うかもしれませんが、実はこれは現実的に（ほぼ）不可能なのです。

たとえば、有意水準を 0 にする最も簡単な方法は、**全て「効かない」と判断**することです。しかし、このとき、「効く」薬でも全てに「効かない」という判断をすることになるので、検出力も 0 に下がってしまいます。逆に、検出力を 1 にする最も簡単な方法は、**全て「効く」と判断**することです。しかし、このとき、「効かない」薬全てに「効く」という判断をすることになりますので、有意水準は 1 に上がってしまいます。

大事なことは、**有意水準を下げれば検出力も下がる、検出力を上げれば有意水準も上がる**ということです。そして、当局からは有意水準が大きくならないよう（大体両側なら 5 %、片側なら 2.5 %にするよう）に要請がある ので、**まず有意水準が決まり、そのあとに検出力を考える**という順番なのです[*3]。では、「決まった有意水準に対して検出力を上げる」にはどうしたらいいのでしょうか？ 実はそこに例数の出番があるのですが、しかしそれをご説明するにはもう少し準備が必要です。

2.3 より正確に考えると

今後のために、正確に考えていきましょう。
有意水準とは、「**データが帰無仮説に従っているにも関わらず、帰無仮説が棄却されてしまう確率**」であり、**検出力**とは「**データが対立仮説に従っている場合に、正しく帰無仮説が棄却される確率**」という風に言われることがよくあります。有意水準についてはこれは正しいのですが、検出力については厳密にはこの表現は間違いです[*4]。この点については、あとから詳しく見ていきます。

3 「標本平均の分布」と検定

3.1 「標本平均の分布」とは

まず**「標本平均の分布」**とは何かを見ていきます。最も重要な点として、我々は基本的に試験は 1 回しか行いません。ですので、1 回の試験で標本平均は 1 つの値しか得られません。では、その**得られた標本平均が信頼できる値か否かはどのように考えればよいのでしょうか**？

[*2] また、規制当局からは「効いているのに効かないと判断される確率が高いということは、効く薬が製品化できない可能性が高いということである。そのような試験に被験者を募って治験薬を投与することは倫理的に問題がある」という観点から、こちらの確率も的確に制御するように要請を受けることが多いです。

[*3] あくまで「考え方」の順番で、実際の試験計画時には同時に決めます。

[*4] 簡単に言いますと、こういう感じです。降圧薬を考えます。主要評価項目はベースラインからの血圧減少量で、片側検定をするとします。このとき、対立仮説は「実薬群 (μ_A) の方がプラセボ群 (μ_P) よりも減少量が**大きい**($\mu_A > \mu_P$)」です。ところが、「実薬群の減少量がプラセボ群に比べて 10 大きい ($\mu_A = \mu_P + 10$)」を検出する検出力と「実薬群の減少量がプラセボ群に比べて 15 大きい ($\mu_A = \mu_P + 15$)」を検出する検出力は異なるのです。

実は、こういう風な発想をするのです。つまり、もし仮に同じ試験をたくさん繰り返していたら、この標本平均値はどのように変わっていくか という発想です。たとえば、日本国民全体の平均血圧を推定しようとするときに, (試験 A)「5 人のデータの標本平均が 130 だった」、(試験 B)「1,000,000 人のデータの標本平均が 130 だった」、とします。このとき (試験 A) は「人数が少なすぎるので、同じ調査を何回も繰り返したら 130 から結構ずれた値もたくさん出てくるに違いない。だから日本国民全体の平均が 130 とはなかなか強く言えない」と思われる方が多いでしょう。一方、(試験 B) は「人数が結構多いので、同じ試験をくり返しても大体 130 に近い値になるに違いない。だから、大体日本国民の平均は 130 くらいと考えてよいのでは？」と思われるでしょう。このように、**仮想的に同じ試験をたくさん繰り返して、得られた値のバラツキ具合から信頼性を考える**のです。

さてそう考えると「分布」のイメージをつかむのは容易です。つまり、**「同じような試験をたくさんくり返して、出てきた標本平均の値のヒストグラム」**を作ります。このヒストグラムを 標本平均の分布 という風に考えていただければ結構です。そして、この 標本平均の分布のばらつきが大きい ときは、「次に同じ試験をしたら、結構値が変わるかもしれない」ということで、**「標本平均の値はそれほど信頼がおけない」**、と判断されます。一方、ばらつきが小さい ときは、「次にやっても大体同じ値になるはず」ということで**「標本平均の値は信頼できる」**と判断されることになります。

3.2 「標本平均の分布」を用いた検定

では次に**「標本平均の分布」**とそれを用いた検定を考えていきます。検定は、とりあえず片側で考えていきます。つまり、帰無仮説と対立仮説として

$$H_0 : \mu = 0$$
$$H_1 : \mu > 0$$

のような状況を頭に入れておいてください。

さて、検定を考えるときに知りたいのは「データ 1 つ 1 つの値がいくつか」ではなくて「平均がいくつか」の方です。平均がいくつかを推定した値が標本平均ですので、検定の主役は「データ（の分布）」ではなくて「標本平均（の分布）」ということになります[*5]。

3.3 「標本平均」の分布

上の状況で、有意水準 2.5 %の片側検定（上側）とは「統計量[*6]を計算し、帰無仮説のもとで統計量の従う分布の確率密度関数を考え、それの上側 2.5 %点より大きい値だった場合に棄却する」という手順をとります。つまり、

(i) 統計量を計算する

(ii) 帰無仮説のもとでの、統計量の確率密度関数を考える

(iii) (ii) の確率密度関数の上側 2.5 %点を計算し、(i) の統計量の値と比較する

の 3 ステップが必要です。

例えば、日本全国の収縮期血圧の分布が $N(120, 20^2)$ だったとします[*7]。このとき、**A 県、B 県の平均がそれぞれ全国の平均と比べて同じかそれとも A 県、B 県の方が高いのか**を知りたかったとします[*8]。

仮説を書いておきます。A 県の平均値を μ_A, B 県の平均値を μ_B とおくと[*9]、

[*5] たとえば 10 人の標本平均が 0.1 のときは「標本平均の分布のばらつき（標準誤差) が大きい」ので $H_0 : \mu = 0$ が棄却できないけれど、10,000 人の標本平均が 0.1 のときは「標本平均の分布のばらつき（標準誤差）が小さい」ので $H_0 : \mu = 0$ が棄却できる、というようなことがありえます。これは、「10 人の標本平均の分布」と「10,000 人の標本平均の分布」が異なるからです。

[*6] 上の例では標本平均です。

[*7] 適当に書いていますので、実際とは異なると思います。ご了承ください。

[*8] 今回は、検定の多重性については一切考えないことにします。

[*9] もう一度注意しておきますが、この平均は「真の値（未知)」です。

A 県の場合は、

$$H_0 : \mu_A = 120 \quad \text{（全国平均と同じ）}$$
$$H_1 : \mu_A > 120 \quad \text{（A 県の方が高い）}$$

であり、B 県では

$$H_0 : \mu_B = 120 \quad \text{（全国平均と同じ）}$$
$$H_1 : \mu_B > 120 \quad \text{（B 県の方が高い）}$$

となります。なお、上の注釈にも書きましたが検定の多重性は本稿では考慮しません。

ここで、

(a) A 県の 200 人のデータ $x_1, \cdots, x_{200}$ の標本平均の分布

(b) B 県の 600 人のデータ $y_1, \cdots, y_{600}$ の標本平均の分布

を考えてみます。**仮にどちらの県も帰無仮説（全国と同じ分布に従う）が正しいとすると、**データの分布 は両方同じになります。つまり、

$$x_1, \cdots, x_{200} \ \sim \ N(120, 20^2)$$
$$y_1, \cdots, y_{600} \ \sim \ N(120, 20^2)$$

です。そして、標本平均の分布 はそれぞれ、

(a) の A 県の 200 人の標本平均 $\bar{x}$ は、平均が 120, 分散が $\frac{20^2}{200} = 2$ の正規分布 $N(120, 2)$ に従い

(b) の B 県の 600 人の標本平均 $\bar{y}$ は、平均が 120, 分散が $\frac{20^2}{600} = \frac{1}{3}$ の正規分布 $N(120, \frac{1}{3})$ に従う

となります。

では各県の**標本平均の分布**を図にしてみましょう。2 つのグラフの軸の尺度は合せてあります。

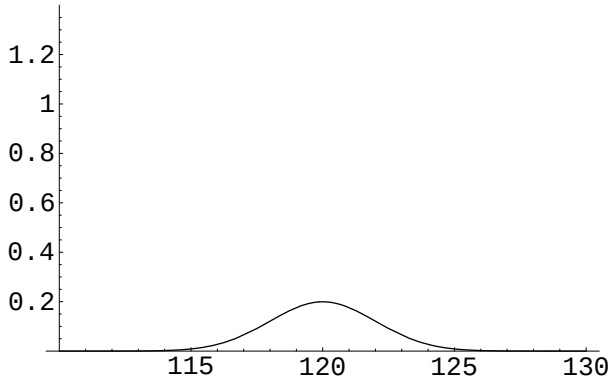

図 1　A 県の 200 人の収縮期血圧の**標本平均の分布**：
$N(120, 2)$ の確率密度関数

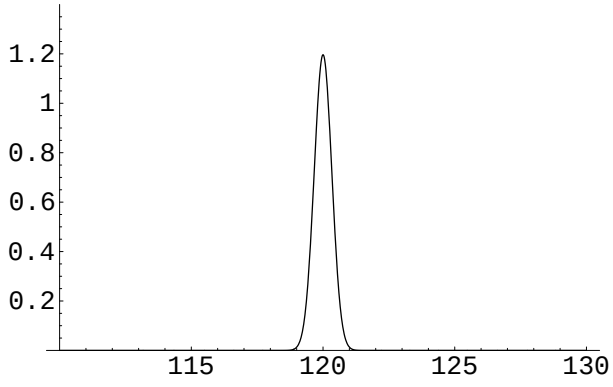

図 2　B 県の 600 人の収縮期血圧の**標本平均の分布**：
$N(120, \frac{1}{3})$ の確率密度関数

このように、**データ自身の分布は同じでも、200 人の標本平均の分布と 600 人の標本平均の分布とでは、分布形が大きく異なっています**。

では、それぞれの分布の上側 2.5 %点を比べてみましょう。この点は、各県の検定の棄却限界です。見やすさを考えて、今度は２つのグラフの軸の尺度を変更してあります。

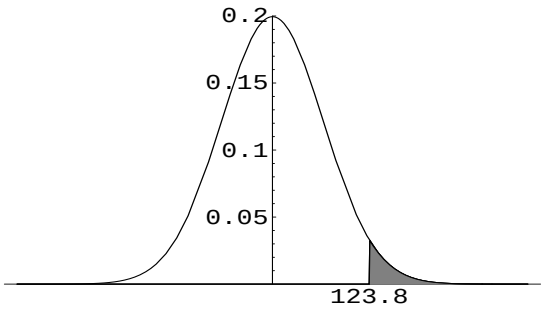

図 3　帰無仮説のもとでの A 県の 200 人の**標本平均の分布** $N(120, 2)$ の上側 2.5 %検定の棄却域

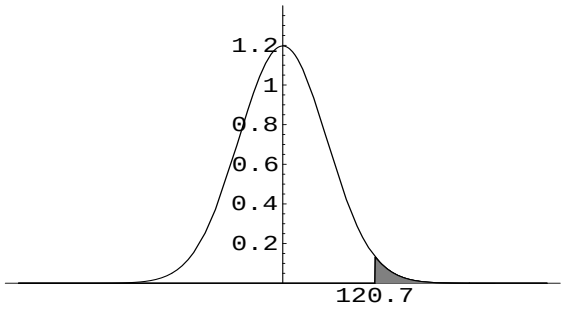

図 4　帰無仮説のもとでの B 県の 600 人の**標本平均の分布** $N(120, \frac{1}{3})$ の上側 2.5 %検定の棄却域

さて、2 つの図を見比べて何が分かるでしょう？

一見して、**上側 2.5 %点を与える数値が異なっている**ことが分ります。これはつまりこういうことです。

「『平均が 120 より大きい』と言いたいときに、200 人の標本平均だったら 123.8 を超えないといけないのに対して、600 人の平標本均だったら 120.7 を超えればよい」ということです[*10]。同じ帰無仮説を棄却したいときに、例数が大きければ値が小さくてよい のです。

つまり、たとえば

・200 人 の標本平均が 122 となった場合

$\Longrightarrow$「真の値が 120 であっても、データのばらつきを考えれば 122 くらいになることはある」と判断される

（帰無仮説が棄却されない）

・600 人 の標本平均が 122 となった場合

$\Longrightarrow$「データのばらつきを考慮しても、真の値が 120 とは考えにくい」と判断される（帰無仮説が棄却される）

となります。このように、**「データ数が多い」ことで、標本平均の値は同じでも「その値の信頼性が高くなっている」**わけです。

これが例数設計の際に非常に重要になってくるポイントです。

4　目で見る有意水準・検出力

以下、検定は有意水順 2.5 %の片側検定を仮定します。

4.1　目で見る有意水準

では、2.3 節においてきちんと表現した有意水準を「目で見て」みましょう。とはいっても、実はもう既に見ているのです。有意水準、つまり**「帰無仮説が正しいにも関わらず、帰無仮説が棄却されてしまう確率」**というのは、図 3, 図 4 の塗りつぶされた部分です。つまり、「有意水準を片側 2.5 %にしなさい」という要請をグラフの言葉で言うならば、「図 3, 図 4 の塗りつぶされた部分の面積が 0.025 になるようにしなさい」という要請と言い換えることができます。

[*10] 棄却限界は有意水準と例数が決まれば検出力とは関係なく決まります。

4.2 目で見る検出力

有意水準を見たので、次は検出力です。しかし、**実は検出力を図に表す前に、検出力を正確に定義する必要があります**。それについて考えていきましょう。いま、帰無仮説・対立仮説は以下のように与えられています。

$$H_0 : \mu = 120$$
$$H_1 : \mu > 120$$

ここで、帰無仮説のもとでの データの分布 は平均 120、分散 20^2 の正規分布なので $N(120, 20^2)$ です。ですから有意水準の場合は、標本平均の分布を求めて $N(120, 2)$ や $N(120, \frac{1}{3})$ のグラフを描けばよかったのです。

ところが、**対立仮説は「平均が 120 より大」というあいまいな与え方をしています。これではデータの分布が一つに決まらないので、グラフが描けない**のです（実際の薬効は試験開始前には分らないので、仮説としてはこうするしかないのですが)。これは大変大きな問題です。しかし「決まらない」と言っても始まりませんので、**とりあえず「えいやっ」と一つ決めてしまいましょう**。例えば、

$$H_1' : \mu = 122$$

としてしまいます (さらに、バラツキは変化しないことも仮定しておきます)。ここで、本稿だけの用語ですが、H_1' を**「見込みの対立仮説」**と呼びます。こうすれば、H_1' のときにデータの従う分布が $N(122, 20^2)$ と一つに決まりますので、めでたく分布を書くことができるようになります。さて、このときに **検出力** とは何かを整理しますと、「**データが $N(122, 20^2)$ に従っているときに、$N(120, 20^2)$ に従っていない、と正しく判断される確率**」です。では、標本平均の従う分布を考えていきましょう。いま、データは $N(122, 20^2)$ に従うと仮定していますので、A 県の 200 人の標本平均の従う分布は、$N(122, 2)$ であり、B 県の 600 人の標本平均の従う分布は $N(120, \frac{1}{3})$ となります。そして有意水準 2.5 %の片側検定なので、図 3・図 4 より、それぞれ 123.8 や 120.7 を超えたときに、帰無仮説を棄却することになります[*11]。

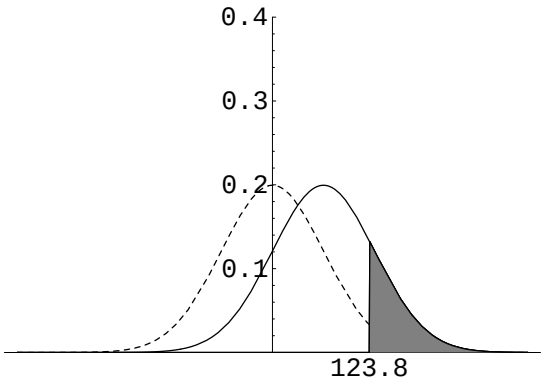

図 5　A 県の 200 人の**標本平均の分布が** $N(122, 2)$ のときの検出力。実線は H_1' が正しいとき、点線は H_0 が正しいときの標本平均の分布。

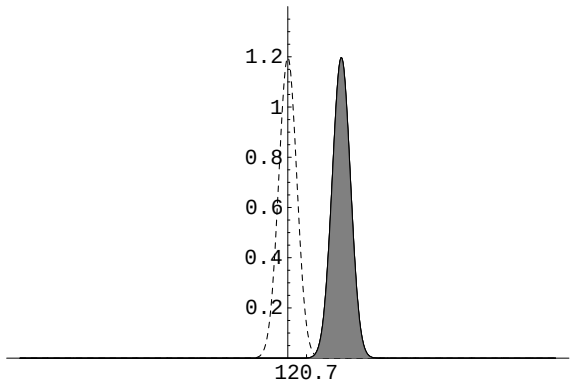

図 6　B 県の 600 人の**標本平均の分布が** $N(122, \frac{1}{3})$ のときの検出力。実線は H_1' が正しいとき、点線は H_0 が正しいときの標本平均の分布。

「検出力」とは 現実が見込みの対立仮説 H_1' に従っているときに、正しく帰無仮説 H_0 を棄却できる確率 であり、現実が対立仮説 H_1 のときではありません。これはつまり、要は**「事前に見込んだ通りの差 があるときに、正しく差があるという判断ができる確率」**です。つまり、例数設計を行う際には、対立仮説だけではなくて「どのくらいの差を見込むか」ということを考えなくてはなりません。

さて、検出力を上の図で言うと、「**仮説 H_1'（実線のグラフ）が正しいときに、帰無仮説 H_0 が棄却される（棄却限界を超える）確率**」なので、塗りつぶされた部分の面積が検出力となります。200 人の平均である図 5 では検出力は 3 割にも満たない程度、600 人の平均である図 6 では検出力はほぼ 1 であることが一見してお分かりいただけるでしょう。**人数が増えると、データの分布が同じで有意水準 (α) も同じでも検出力が増加する**ことがお分かりいただけましたでしょうか。

[*11] 棄却限界は常に「有意水準と例数の 2 つ」から決定されることに注意してください。

4.3　重要な注意：医学的に意味のある差

今までの話から、「例数を増やせば試験は検出力が増えて試験は成功しやすくなる」ということはご理解いただけたと思います。では、「例数が増えれば差が出やすくなってよいことしかない」なのでしょうか？

実はよくないことが起こってしまう可能性があるのです。というのは、先の図 5・図 6 から今回の例で 600 人の標本平均で考えた場合、「実際の収縮期血圧の平均値が 120 より 2 しか大きくない」場合でさえ、ほぼ検出力が 1 となってしまいます。さらにもっと例数を増やしてたとえば 1,000,000 例くらい集めますと、**「実際の収縮期血圧の平均値が 120 より 0.1 だけ大きい場合」でさえ、ほぼ検出力が 1 になってしまいます**。つまり、例数が多過ぎるために、たった 0.1 の違いでも敏感に検出して「平均値は 120 より大きいですよ」という結論を出してしまうのです。これでは、検定の結果が医学的に意味を持たなくなってしまいます。

そのため、「正しい例数を設計する」ことが非常に重要になってきます。具体的には、**「医学的に意味のある差」**を先に決めるのです（これには医学的知識や類薬の情報、前の試験の情報などを利用します)。その値を Δ（実際は数字）とすると、たとえば**「プラセボ群よりも平均値が Δ だけ大きいというのは医学的に意味があるので、そのとき 80 %は検出できるように（＝検出力を 0.8 に）しましょう」**という形で例数設計を行うのです（つまり、先の例でしたら $H_1' : \mu = 120 + \Delta$ とするわけです)。製薬では、この「医学的意味のある差」のことを「期待される薬効」などと言うこともあります。

5　例数設計のやり方

これで準備は整いました。では、例数設計のやり方に入りましょう。

5.1　例数設計に必要なもの

まず、最初に指定すべきは

(i)　有意水準：α

(ii)　検出力：$1 - \beta$

の 2 つの値です。さらに、4.3 節で述べたように、

(iii)　医学的に意味のある差 Δ（期待される薬効）

が必要です。そしてさらに、先ほどはさらっと流してしまいましたが、

(iv) データの分散 σ^2

も、簡単のため今回は既知としましょう。分散が既知、というのは「前の試験のデータの推定値を参考に決める」という意味だと考えてください。

5.2　例数設計のやり方（数値例）

「例題 1」

降圧薬 A とプラセボを比較する臨床試験を計画したいとします。各群の血圧の減少量のデータはそれぞれ、分散 400（標準偏差 20）の正規分布に従うことが分かっているものとします。さらに降圧薬 A は、プラセボと比較して平均して収縮期血圧を 10 下げることが見込まれている とします ($\Delta = 10$)。このとき、このとき、有意水準 2.5 %、検出力 80 %の片側検定を行うのに必要な例数を計算してください。

「考え方 : 例題 1」

まず、（当然）例数が分からないので、1 群あたり n 人 だとしておきましょう。このとき、実薬群の収縮期血圧の減少量（を表す確率変数）を $X_1, \cdots, X_n$ とし、プラセボ群の収縮期血圧の減少量（を表す確率変数）を $Y_1, \cdots, Y_n$ とします。両群とも、データは分散 400（標準偏差 20）の正規分布に従うことが分かっていますので、プラセボ群の収縮期血圧の減少量

の従う分布を $N(\mu_y, 400)$, 実薬群の収縮期血圧の減少量の従う分布を $N(\mu_x, 400)$ と書くことにします。

ここで、帰無仮説・対立仮説は

$$H_0 : \mu_y = \mu_x$$
$$H_1 : \mu_y < \mu_x$$

と書けます。さらにいま、$\Delta = 10$ を見込んでいるので、見込みの対立仮説を、

$$H_1' : \mu_x = \mu_y + 10$$

とおきましょう。

書き直すと、

$$H_0 : \mu_x - \mu_y = 0$$
$$H_1 : \mu_x - \mu_y > 0$$
$$H_1' : \mu_x - \mu_y = 10$$

と書けます。

さて、両群の平均値の差に興味があるので、まずそれぞれの平均値を $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ とおきます。このとき、各標本平均の従う分布は $\overline{Y} \sim N\left(\mu_y, \frac{400}{n}\right)$, $\overline{X} \sim N\left(\mu_x, \frac{400}{n}\right)$ です。これより、標本平均の差

$$d = \overline{X} - \overline{Y}$$

の従う分布を考えましょう。すると、正規分布の性質[*12]より、

$$d \ \sim \ N\left(\mu_x - \mu_y, \frac{800}{n}\right)$$

となります。この 統計量 d が帰無仮説 $H_0 : \mu_x = \mu_y$ と見込みの対立仮説 $H_1' : \mu_x = \mu_y + 10$ のもとで従う分布をそれぞれ考える のです。

H_0 に従うとき $\mu_x - \mu_y = 0$ より

$$d \ \sim \ N\left(0, \frac{800}{n}\right)$$

また、H_1' に従うとき $\mu_x - \mu_y = 10$ より

$$d \ \sim \ N\left(10, \frac{800}{n}\right)$$

となります。

これが大体、以下の図 7 のようになればよいわけです。

[*12] 一般に $X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$ とおくと、$X - Y \sim N(\mu_x - \mu_y, \sigma_x^2 + \sigma_y^2)$ です。

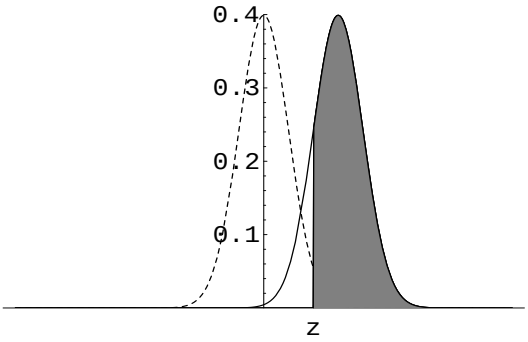

図 7　有意水準 2.5 %、検出力 80 %のときの標本平均の分布（点線が H_0 が正しいとき、実線が H_1' が正しいとき）

さて、このとき図 7 中の**棄却限界 z を、2 つの立場で考えます**。

(a) 帰無仮説 H_0 の立場

帰無仮説 H_0 の立場（図 7 の点線）では、z は有意水準 0.025 の棄却限界です。d の従う分布は $N\left(0, \frac{800}{n}\right)$ でしたので、z は $N\left(0, \frac{800}{n}\right)$ の上側 2.5 %点（下側 97.5 %点）となります。標準正規分布と結び付けると、

$$\frac{z-0}{\sqrt{\frac{800}{n}}} = z_{0.975} \tag{1}$$

です[*13]。「$z_{0.975}$=「標準正規分布の上側 2.5 %点」　1.96」を用いつつ、z が主役になるように整理してやると、

$$z = 1.96 \cdot \sqrt{\frac{800}{n}} \tag{2}$$

となります。

(b) 見込みの対立仮説 H_1' の立場

H_1' の立場（図 7 の実線）では、**検出力 0.8 より z より右に全体の面積の 80 %があります**。H_1' のもとで d の従う分布は $N\left(10, \frac{800}{n}\right)$ でしたので、z は $N\left(10, \frac{800}{n}\right)$ の下側 20 %点 となります。(a) と同じく標準正規分布に結び付けると、

$$\frac{z-10}{\sqrt{\frac{800}{n}}} = z_{0.20}$$

となります。整理すると「$(z_{0.20} = -0.84)$」から、

$$z - 10 = -0.84 \cdot \sqrt{\frac{800}{n}}$$
$$z = 10 - 0.84\sqrt{\frac{800}{n}} \tag{3}$$

となります。

[*13] 一般に、$N(\mu, \sigma^2)$ の下側 $(100 \cdot \alpha)$ %点を z とおくと、z と標準正規分布 $N(0,1)$ の下側 $(100 \cdot \alpha)$ %点 z_α との関係は

$$\frac{z-\mu}{\sqrt{\sigma^2}} = z_\alpha \iff z = \mu + z_\alpha\sqrt{\sigma^2}$$

となります。

ここで、(2) と (3) は同じ z が出てきています。この z **は「この検定の棄却限界」という全く同じものですので消去して計算します。**すると、

$$1.96\sqrt{\frac{800}{n}} = 10 - 0.84\sqrt{\frac{800}{n}}$$
$$1.96\sqrt{\frac{800}{n}} + 0.84\sqrt{\frac{800}{n}} = 10$$
$$2.8\sqrt{\frac{800}{n}} = 10$$

となります。次に、両辺に $\sqrt{n}$ をかけると、

$$2.8\sqrt{800} = 10\sqrt{n}$$
$$\sqrt{n} = \frac{2.8\sqrt{800}}{10}$$

となり、さらに両辺 2 乗すると、

$$n = 62.72$$

となります。以上より、まあ小数のところは多目に見積もって**「1 群あたり 63 例」**という結果になります[*14]。

5.3 例数設計のやり方 (式の計算)

では一般論として、先の例題の数字だったところを文字にしてやってみましょう。

「例題 1'」

降圧薬 A とプラセボを比較する臨床試験を計画したいとします。各群の血圧の減少量のデータはそれぞれ、分散 σ^2 (両群で共通) の正規分布に従うことが分かっているものとします。さらに降圧薬 A は、プラセボと比べて 平均して収縮期血圧を Δ 下げることが見込まれている とします。このとき、このとき、有意水準 α、検出力 β の片側検定を行うのに必要な例数を計算してください。

「考え方 : 例題 1'」

数値例と同じように考えていきます。例数を n として、実薬群の収縮期血圧の減少量を表す確率変数を $X_1, \cdots, X_n$ とし、プラセボ群の収縮期血圧の減少量を表す確率変数を $Y_1, \cdots Y_n$ とします。ここで、$X_1, \cdots, X_n \sim N(\mu_A, \sigma^2)$、$Y_1, \cdots Y_n \sim N(\mu_P, \sigma^2)$ とします。帰無仮説と対立仮説は

$$H_0 : \mu_P = \mu_A$$
$$H_1 : \mu_P < \mu_A$$

と書けます。今、降圧薬 A では Δ 収縮期血圧が下がることを見込んでいるので、見込みの対立仮説は

$$H_1' : \mu_A = \mu_P + \Delta$$

とおきます。整理すると、

$$H_0 : \mu_A - \mu_P = 0$$
$$H_1 : \mu_A - \mu_P > 0$$
$$H_1' : \mu_A - \mu_P = \Delta$$

となります。

[*14] 今回は簡単のため「脱落 0 %」を想定しています。現実的には、脱落率などを考えてもう少し増やすことになると思います。

ここで、まず $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ とおきます。このとき、$\overline{Y} \sim N\left(\mu_P, \frac{\sigma^2}{n}\right)$, $\overline{X} \sim N\left(\mu_A, \frac{\sigma^2}{n}\right)$ となります。ここで、標本平均の差

$$d = \overline{X} - \overline{Y}$$

の従う分布を考えると $d \sim N\left(\mu_A - \mu_P, \frac{2\sigma^2}{n}\right)$ となります。これは、帰無仮説 $H_0 : \mu_A - \mu_P = 0$ のもとでは、

$$d \sim N\left(0, \frac{2\sigma^2}{n}\right)$$

となり、見込みの対立仮説 $H_1' : \mu_A - \mu_P = \Delta$ のもとでは、

$$d \sim N\left(\Delta, \frac{2\sigma^2}{n}\right)$$

となります。

以下、帰無仮説・対立仮説のそれぞれの立場で考えましょう。

(a) 帰無仮説 H_0 の立場

帰無仮説の立場では、先の z は $N\left(0, \frac{2\sigma^2}{n}\right)$ の上側 $(100 \cdot \alpha)$ %点（つまり、下側 $100 \cdot (1-\alpha)$ %点）です。従って、標準正規分布に直すと

$$\frac{z-0}{\sqrt{\frac{2\sigma^2}{n}}} = z_{1-\alpha}$$

$$z = z_{1-\alpha} \cdot \sqrt{\frac{2\sigma^2}{n}}$$

$$z = z_{1-\alpha} \cdot \sqrt{\frac{2\sigma^2}{n}} \tag{4}$$

となります。

(b) 見込みの対立仮説 H_1' の立場

見込みの対立仮説 H_1' の立場では、z は、$N(\Delta, \frac{2\sigma^2}{n})$ の下側 $(100 \cdot \beta)$ %点 です。標準正規分布に直すと、

$$\frac{z-\Delta}{\sqrt{\frac{2\sigma^2}{n}}} = z_{\beta}$$

$$z - \Delta = z_{\beta} \cdot \sqrt{\frac{2\sigma^2}{n}}$$

$$z = \Delta + z_{\beta} \cdot \sqrt{\frac{2\sigma^2}{n}} \tag{5}$$

となります。

(4) と (5) より z **を消去すると、**

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

となり、両辺に $\sqrt{n}$ をかけると、

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

$$\Delta\sqrt{n} = z_{1-\alpha} \cdot \sqrt{2\sigma^2} - z_{\beta} \cdot \sqrt{2\sigma^2}$$

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

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

となります。次に、両辺 2 乗すると、

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

です。さらに、正規分布の左右対称性から $z_{1-\alpha} = -z_\alpha$ を代入すると、

$$n = \frac{2\sigma^2(-z_\alpha - z_\beta)^2}{\Delta^2} \tag{6}$$

$$= \frac{2\sigma^2(z_\alpha + z_\beta)^2}{\Delta^2} \tag{7}$$

となります。**この n が、片側検定の場合の平均値の比較における例数**になります。

5.4 両側検定の場合

最後に一瞬だけ両側検定についても触れましょう。有意水準 α のとき、それを両側に $\frac{\alpha}{2}$ ずつ振り分けるので、うるさいことを抜きにすると、(7) の α を $\frac{\alpha}{2}$ に置き換えた、

$$n = \frac{2\sigma^2(z_{\frac{\alpha}{2}} + z_\beta)^2}{\Delta^2}$$

で大体の値が求まります[*15]。

[*15] もう少しだけ言いますと、両側検定のために出てくるもう一方の側は「無視できるくらい確率が小さいので無視」するのです。そうすると実質片側検定と同じと考えられます。

6　終わりに

まとめましょう。平均値の差に関する例数設計で、**両群ともにデータが正規分布に従い、両群の分散が等しいことは仮定します。**

例数設計に必要なもの

有意水準 (α)、検出力 ($1-\beta$)、見込まれる薬効 (Δ)、データの分散 (σ^2)

求め方の手順

(i)　両群のデータの従う分布を書く。

(ii)　標本平均の差 d の従う分布を書く。

(iii)　帰無仮説・見込みの対立仮説をきちんと書き、それぞれの場合に差 d の従う分布がどのようになるかを見る。

(iv)　帰無仮説・見込みの対立仮説のそれぞれが正しい場合の標本平均の従う分布の図を描いて、「有意水準 α, 検出力 $1-\beta$」が目に見えるようにする。

(v)　棄却限界点を z とおく。

(vi)　z を、帰無仮説・対立仮説それぞれの立場で意味づけし、標準正規分布のパーセント点 ($z_{1-\alpha}, z_{\beta}$) で表す。

(vii)　(vi) の 2 式を $z=$ の形に表し、z を消去して、$n=$ の形に直す。

公式

(片側検定)

$$n=\frac{2\sigma^2(z_{\alpha}+z_{\beta})^2}{\Delta^2}$$

(両側検定)

$$n=\frac{2\sigma^2(z_{\frac{\alpha}{2}}+z_{\beta})^2}{\Delta^2}$$

(最後に注意)

n は「1 群あたりの人数」であることを忘れないでください。

7　補足：SAS による実行

7.1　プログラムと出力

最後に、SAS の proc power でこの例数設計を行うとどうなるかを見ておきます。設定は**「例題 1」**とほぼ同じ

・両群の差 : $\Delta = 10$
・標準偏差 : 10, 15, 20 の 3 通り
・有意水準 : 両側 5 %
・検出力 : 80 %

とします。

なお、**「例題 1」**では、分散を既知と仮定して正規分布による検定を用いましたが、今回はより正確に t 検定を用います[*16]。このとき、プログラムは

```
proc   power;
   twosamplemeans   test  =  diff
   meandiff  =  10
   stddev  =  10,  15,  20
   alpha  =  0.05
   power  =  0.8
   ntotal  =  .  ;
run;
```

となります。**twosamplemeans から ntotal までセミコロン (;) がありません**ので注意してください。

では、出力の主要な部分を見てみることにします。

Two-sample t Test for Mean Difference

Computed N Total			
Index	**Std Dev**	**Actual Power**	**N Total**
1	10	0.807	34
2	15	0.808	74
3	20	0.801	128

となります。

先に指定しました、「標準偏差 (Std Dev)」が 10, 15, 20 の 3 通りが出力されています。「Actual Power（実際の検出力）」とは、例数は 1 ずつしか増えませんので、「厳密に検出力 80 %」とはできないことも多く、「大体 80 %になるように設計しましたけど、厳密にはこうなりました」という現実的な検出力のことです。

また、**「N Total」は両群合わせた例数**です。式の計算でご説明したのは 1 群あたりの例数 でしたので、この違いにはご注意ください。

[*16] 以下を見ていただくとお分かりの通り、結果はほとんど変わりませんが。

7.2 式の計算との違い

さて、数式で計算した**「例題 1」**では、標準偏差 20 で 1 群 63 例となりました。つまり、2 群合わせて $63 \times 2 = 126$ となります。一方、上の SAS の出力では 128 例となり、全体で 2 例ほど異なります。この違いは、

- **「例題 1」**では有意水準 片側 2.5 % だが、今回は 両側 5 %
- **「例題 1」**では分散既知の 正規分布による検定 を行ったが、今回は t 検定

という違いからくるものです。ただ、異なるといってもこの程度で大差はありません。