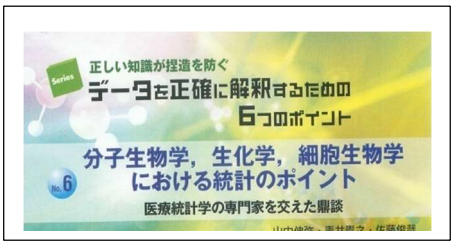
【統計一口メモ　第１２話】

＜外れ値をかってに外してはいけないって言われるけど＞

名古屋市立大学大学院医学研究科 非常勤講師　薬学博士　松本一彦

実務研究者や実務臨床家はデータをとるのが仕事です。でも、そのデータにはしばしばとんでもない値が出ることがあります。そのデータを使うか外すか、みなさんはその基準を持っていますか？「これ外したら有意になるから」、「異常値だもん外さなきゃ―」、こんな答えをすると統計家だけでなく誰もが「かってに外してはダメ！」と言います。そうなんです、かってに外すから怒られるのです。ちゃんと外す“理由”と“基準”があれば誰も文句は言わないのでは？ その基準って何でしょう。まず、外す理由です。「間違えたデータが混入した」とか「試薬を入れ忘れたミス」など理由が明らかなときは、外すか外さないか考える前に捨てていますね。問題は「何でこんな値が出たんだろう？」とわからないときに迷います。今回は、そのような場合に「外れ値」をどうするか、外した後にどう考察するかを考えてみましょう。

§１．統計家の考え方



蛋白質　核酸　酵素　Vol.54 No.13(2009)

山中伸弥先生は統計家に次のような考えをぶつけました1)。「実験データの中で、他とはかけ離れた値を示すもの，つまり外れ値にはいつも頭を痛めます。ある測定値が人的なサンプル採取のミスと考えた場合それを除外したら有意差がでたとしても，除外した後の解析だけを示すのは，不適切に思います。一方で，除外した値は明らかにおかしくて，それを除けば知りたいことの真の答えがわかると考えられるときは両方を示したらよいような気がするのですが」。その答えに、統計家の佐藤俊哉先生は「基本的には，誤りであることのはっきりした証拠がないかぎり省くべきではない。あるいは，身長が 100 メートルとかいうありえない値は削除するしかないですけれども，測定のまちがいが記録に戻ってわかったとか，測定者に聞いてわかったとか，そういうことがないかぎり，それは削除してはいけないと学生にも教えています。いわれたように，そのデータを含めた結果と，除外した結果との両方を示すということですが，それも，両方の結果が同じだったら実験結果を支持することになりますが，違ったらやはり，実験自体の信頼性が疑われることになります」と答えました。佐藤先生は”両方を示すこともダメ“と言います。すなわち、”外すことはまかりならぬ“という意見です。※ボクのつぶやき：身長１００メートルはありえないけど２ｍ３０ｃｍだったら？実務者はそこで悩むんだよね。やっぱ、何か物差しが欲しいなー。

海外の統計家の意見 2)：大半のデータと極めて乖離した観測値であっても、その確実性を疑わせるような特別な理由が無いのであるなら、解析から除外

すべきではない。こういった外れ値を除外したのであるなら、いずれの場合も報告しなければならない。※ボクのつぶやき：外してもいいってことじゃん。

外れ値は統計解析に著しい影響を与える場合があるので、こういった値に結論がどの程度左右されるかを評価するために、その観測値を含む場合と含まない場合の双方のデータを解析することは有益である。※ボクのつぶやき：Dr. Altman は、両方の結果を示すのは OK というスタンスをとっているんだ。

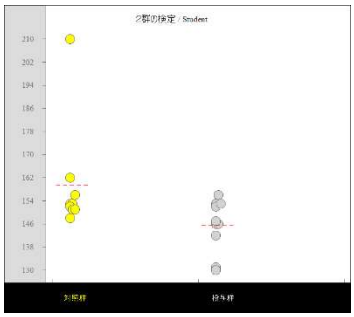
§２．外れ値を確認するための「箱ひげ図」

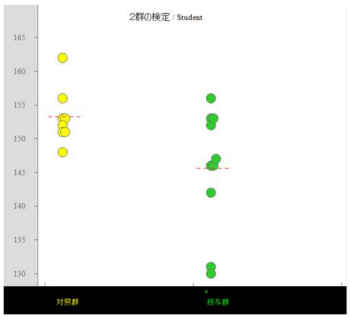
＜ｔ検定で対照群と投与群の間で有意差がなかったのですが、そのまま諦めますか？＞

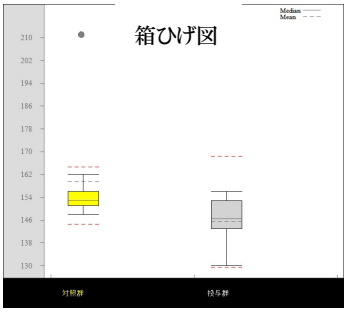
対照群に１例、外れ値らしき数値があります。これを外すと、結果は有意となりました。

実務者は悩みます。そこで、“箱ひげ図”で確認したところしっかり外れ値ということが分かりました。そこで、外した後の数値を論文に書いて、考察に「210」を外したことを明記して事なきを得ました。

この箱ひげ図とはいったい何ものでしょう。







＜箱ひげ図＞

外れ値をデータ解析に含めるか否かを検討し、適正な解析に臨むためのツールです。

理論的には 100 例中１例（１％）をみつけることができます。正規分布に近いかどうか（ひげの長さ

等しいか)、バラツキが大きいかどうか（箱の大きさ）も目で確かめることができます。

箱の長さは四分位範囲といって観測値の 50%が入る幅です。その 1.5 倍のところが限界線で、　　そこからはみ出たものが外れ値です。

§３．外れ値の取り扱い

外れ値は生物実験データには必ず存在し、その検討は避けて通れない。

- データを見たときには、平均値だけではなく、必ず生データから散布図を作成し、箱ひげ図により、外れ値の有無を確かめる。
- 平均値と中央値が求められていて、生データがないときは 2 者の食い違いから外れ値の存在を疑う。Pharmaco は箱ひげ図の中に平均値と中央値を表示します。
- 平均値と標準偏差が求められていて、生データがないときは標準偏差の大きさから外れ値の存在を疑う。自然界の標準偏差は平均値の 15～20%といわれています。それ以上のときは、バラツキの大きい原因に外れ値を疑うことも必要です。

§４．外れ値への対応（まとめ）

- 外れ値の原因が測定ミスの場合：明記して棄却。
- 測定ミスとは断定できないが、生物学的にはかなり異常：外れ値を入れた場合と除いた場合とで比較考察する。
- 外れ値を決定樹にしたがってスミルノフ・グラブス検定などで自動的に外して解析することは、厳に慎むべきである。外れ値はデータを得た人が責任をもって処理しなければならない。

§５．Pharmaco での解析

図には、限界線と平均値、中央値が表示されます。

表には、上側と下側の四分位値が表示されます。

文献

1） 山中伸弥他． 蛋白質核酸酵素 54 巻 13 号 2009 年

2） M．J．Gardner，D．G．Altman 「信頼性の統計学」 サイエンティスト社 2001 年