#017★★★☆ 2021.7.19 / 2023.7.25

外れ値の確認方法とその扱いについて。

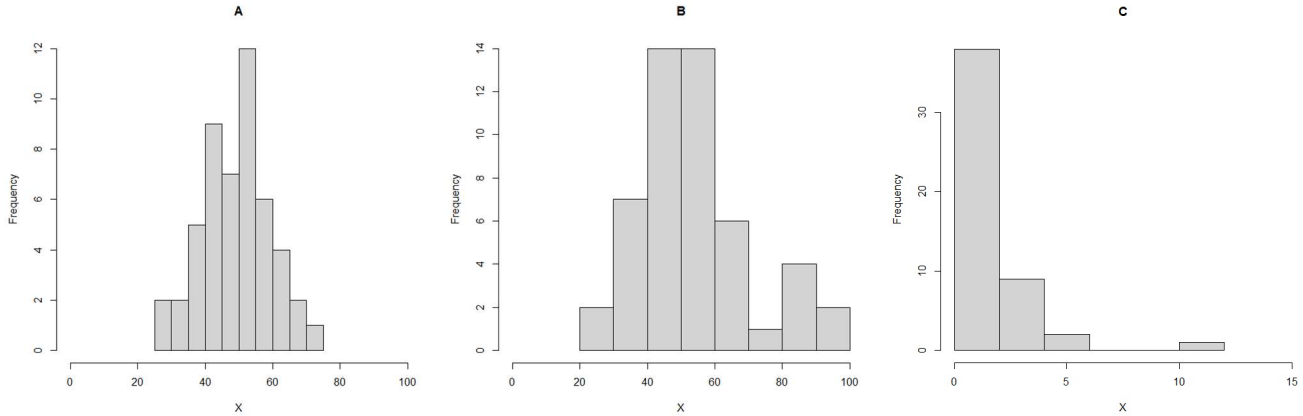
検定や回帰分析といったデータ解析では、常にデータの分布を意識して適切な手法を選ぶ必要がある。「外れ値」、すなわちデータ全体の分布からは離れていそうな数値については、その理由が実験（測定）のミスなのか、あるいは誤差として起こり得る値なのかを判断した上で採否を決める。外れ値も解析に含める場合には、データに分布を仮定しないマンホイットニー検定やウィルコクソン順位検定といったノンパラメトリック検定を用いるといった工夫が必要になる。

データの分布を確認した上で 2 群間の平均値比較を行う場合一般には次の表が目安となる。このあと述べるデータセットのうち、A の場合であればパラメトリック検定が、B ではノンパラメトリック検定が、C では対数変換したデータが正規分布に従うことを仮定するなら変換したデータを用いてパラメトリック検定を用いる、という手順になる。

| | パラメトリック検定 | ノンパラメトリック検定 |
|---|---|---|
| 独立 2 標本の検定 | Student-t 検定、Welch-t 検定＊ | Mann-Whitney 検定、Wilcoxon 検定 |
| 対応のある 2 標本の検定 | Paired-t 検定 | Wilcoxon 符号付順位検定 |

＊もとの 2 群間の分散が異なるかどうかで使い分けるとされている。

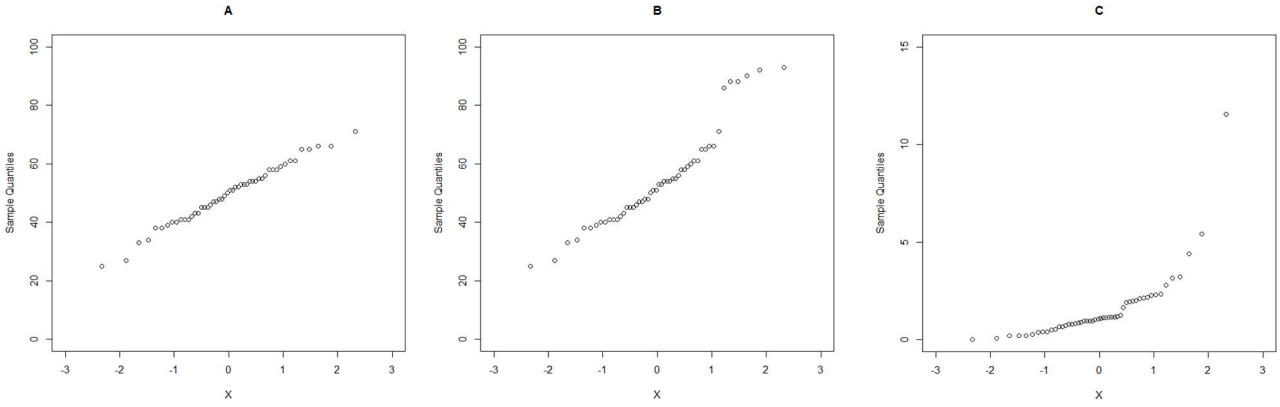
「自身で得たデータの一部が外れ値であり解析から除外するべきなのか、あるいはその誤差はばらつきの範囲内なので解析に含めるべきなのか」についてデータが正規分布に従うかどうかを確認してみる。正規分布は平均値$\mu$と標準偏差$\sigma$（分散$\sigma^2$）とで表される。例としてそれぞれ 50 個のデータから成るデータセット A、B、C を用いて作成し、R 言語（コマンド例：hist(data, main="A", xlab="X", xlim =c(0,100))）で描いたヒストグラムを図に示す。ヒストグラムを見る限り、A、B は正規分布のように見えるが B は 80 以上の数値に「山」がありこれらが正規分布から外れているようにも見える。C は正規分布よりは対数正規分布にも見えるが、10 以上の数値が外れ値かもしれない。




次に、視覚的に正規性を見るための「正規確率プロット」を描画してみる。正規確率プロットのグラフにはいくつかの描画方法があるが、「標準化」「分位点（quantile）」という用語が出てくる。「（正規）標準

化」とは得られたデータを「平均値 0、標準偏差 1 の正規分布（標準正規分布という）」になるように変換することをいい、具体的には元のデータ X について $Y=(X-\mu)\div\sigma$ で計算する。またデータを小さい（あるいは大きい）順に並べたときに各データが全体の何%の位置にあるかを「分位点」という。特に 25%や 75%の点を四分位点といい、また 50%分位点は中央値にあたる。ボックスプロットで汎用される。

正規確率プロットでは実測値と分位点との関係を図示し、正規分布であれば直線関係が得られることを利用する。例えば R 言語で「Q-Q プロット」を作図した場合には、横軸に実測値を標準化した値を（実測値そのままでもグラフの形は同じなので実測値をプロットする方法もある）、縦軸に実測値が正規分布に従うとしたときの各データの分位点をプロットする（コマンド例：qqnorm(data, main="A", xlab="X", xlim=c(-3,3), ylim=c(0,100))）。図のように、A ではほぼ直線関係が得られているが、B では数値の大きい側の 6 ポイントが直線からは外れているようであり、C では明らかに直線関係は見えない。




正規性の検定方法としてシャピロ・ウィルク検定という方法がある（R コマンド例：shapiro.test(data)）。帰無仮説は「データは正規分布に従う」である。A、B、C それぞれで検定を行ってみると p 値はそれぞれ、A：0.920、B：0.00243、C：<0.001 となり、A は帰無仮説を採択（正規分布に従う）、B、C では帰無仮説は棄却（p<0.05）され正規分布に従うとはいえないという結果が得られた。

外れ値の検定方法のひとつとしてスミルノフ・グラブス検定という方法がある。R 言語でこの検定を行う際には左のようなコマンドでライブラリをインストールする必要がある。検定を行ってみると（R コマンド例：grubbs.test(data)）、“p = 0.4023, alternative hypothesis: highest value 93 is an outlier”という結果が表示され、「最大値 93 が外れ値であるとの対立仮説のもので p 値は 0.4023」、即ち 93 というデータはこのデータセットでは外れ値とはいえない。

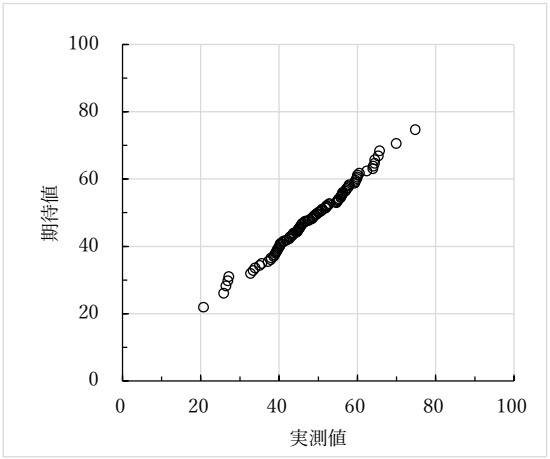
```
> install.packages("outliers")
> library(outliers)
```

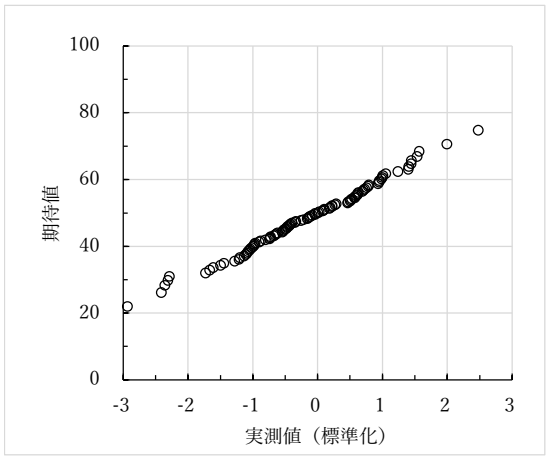
以上のとおり、データが正規分布に従うかどうかは正規確率プロットや正規性の検定で判断でき、またデータが外れ値かどうかを検定する方法もある。しかしながら、実験やカルテ調査でデータを得るのは私たち自身であり、統計的確認をしながらも最終的には私たち自身が適切な判断を行うべきである。操作にミスは無かったか、原液が混合しなかったか、溶血などのミスは無かったか、外れ値らしい数値を示した患者の背景や特徴を精査し臨床的な検証を行ったか、などに注意する。そもそもデータ数が少ない場合には統計解析をしても適切な結論が得られない場合もある。外れ値としてデータを除外するときと除外しないときとでの結果の違いを検証した上で妥当な結論を導くことも重要である。今回示した解析手法は p 値などの客観的事実を得るために必要な手法ではあるが、最後は機械ではなく人間の判断を重視するべきであり、類似の研究領域の先行論文を確認し先行研究が用いている手法を知る姿勢も大切である。

平均値を$\mu$、標準偏差を$\sigma$（分散は$\sigma^2$となる）としたときの正規分布、及び標準正規分布の場合それぞれ右式となる。いくつかの正規確率プロットの描画方法について、その違いを検討した。まず平均値が 50、標準偏差が 10 となる架空の 100 個の数値を、乱数を用いて生成し、それぞれのデータの小さいほうからの順位、全体を 100%としたときの確率（分位点）、さらに正規分布を仮定したときのその分位での値（期待値）を算出した。ここでは Web サイト（https://bellcurve.jp/statistics/blog/15362.html）の記事を参考にした。いくつかの組み合わせでプロットした図を示す。

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \qquad f(x)=\frac{1}{\sqrt{2\pi}}exp\left(-\frac{x^2}{2}\right)$$

左図は（横軸、縦軸）が（実測値、期待値）で、右図は（標準化した実測値、期待値）の図である。横軸を標準化することは横軸の平行移動あるいは縮小（拡大）操作だけなのでグラフの形状そのものは変わらず、図 1、図 2 ともにほぼ直線状であり正規分布が仮定できる。




次の左図は（実測値、確率（分位点））の関係を、右図は（実測値、標準化した期待値）の図で、左図は累積確率プロットに相当し、これは正規分布でも必ずしも直線にならない。右図の縦軸は期待値を標準化した値、即ち標準正規分布の標準偏差の値（±1SD、± 2 SD・・）に相当し、直線状であり正規分布が仮定できる。縦軸を右端図のように累積確率の数値そのもので示した正規確率用紙もあるが、考え方は同じである。