1限目 データの種類

連続データ、カテゴリーデータ

この章の目標

1. mean（平均値）とmedian（中央値）の違いを説明できる。
2. SD（標準偏差）とSEM（標準誤差）の違いを説明できる。
3. 体格指数（BMI）のmean 25、SD 2.4のときに、BMIが26より大きい人は全体の何%くらいいるかを計算できる。

●はじめに

統計がよくわからない、という話をよく聞きます。臨床統計を行っている場合のこの「よくわからない」というのは、統計の難しい計算式を理解できない、というものではなく、

「どのような解析法を使用すればよいのかがわからない」

「その結果の解釈がわからない」

ということだと私は思います。

そもそも統計をどのように行っていくかについて、大きく2つがあります。

Descriptive statistics
Inferential statistics

descriptive statistics は、データを計算し、要約し、データを Table や Figure で示すものです。

もう一つの inferential statistics は、ある集団からのデータを「population」と呼ばれる大きなグループに一般化します。

具体的に見ていきましょう。

ある集団の男女の血圧を測定しました。

	Male	Female
Mean（平均）	123	118
SD（標準偏差）	7	6
Median（中央値）	117	114
Range（範囲）	104-145	102-135

このように、

平均血圧は、男性では平均 123、SD＝7、女性で 118、SD＝6。

中央値は、男性では中央値 117、範囲は 104-145、女性で 114、範囲は 102-135

というようにデータをまとめた状態を descriptive statistics といいます。

これだけみると男性のほうが女性より血圧は高いという風にみることができますね。

しかしここで疑問に思うのは、このデータに基づき、（一般に）男性の血圧は女性より高いといえるか、と言うことです。

これを検討することが Inferential statistics になります。

1-1 統計に出てくる言葉を理解しよう

統計が苦手、という方の中には、用語が難しい、と感じる方もいるのではないでしょうか。英語表記を無理やり日本語にしたような表現もあり、さらに理解を困難にしている気もします。しかしながらそれ以上にわかりやすく端的な言葉がないので、そのような表現になるのは仕方がない部分もあります。

私自身は漢字で理解するほうが堅苦しく難しく感じていますので、基本的には英語のほうで覚えるようにしています。これは英語論文を読む際は英語記載ですし、使用している解析ソフトも英語版であるので、日本語を覚える必要がない、ということにもよります。いずれ英語で覚えないといけないのならば、初めから英語で覚えておいたほうが二度手間にならずよいのではないでしょうか？

それでは統計で出てくる用語を見てみましょう。

●Variable（変数）

あるグループにおいていろいろな値を測定したり、情報を集めると思います。それぞれの項目、値のことを variable（変数）と言います。

例えば、ある集団のデータを集めた場合の、年齢や性別、身長などのそれぞれのことを言います。

Variable には 0 か 1 で表されるような値をとる場合（質問での yes、no、生存や死亡など）と、身長などのように 160 や 166 cm と表現するものがあります。

●Variable のタイプ

それでは variable にはどのようなタイプがあるのか見てみましょう。

大きくは Qualitative/categorical variable と Numerical variable に分けることができます。

■Qualitative/categorical variable

Nominal variable

人種や性別のことです。

これには順序関係はなく、0 を女性、1 を男性、とした場合です。その中間の 1.5 という表現はありません。また、1＝女性、2＝男性と入力することもありますが、2 は 1 の 2 倍、という考えにもなりません。つまり yes、no に近いでしょうか。

Ordinal variable

順序立てされた variable です。例えば、がんのステージ分類では stage 1、stage 2、stage 3、stage 4 とあります。しかしながら stage 2 は stage 1 の 2 倍、stage 4 は stage 2 の 2 倍、とはなりません。このように 1. 2. 3. 4…と並ぶけれども、それぞれの距離が一定でないものとなります。

■Numerical variable

Discrete variable

子供の数のようにある程度の上限があるものです。

例えば統計解析で出産した子供の数が 1 人増えるごとにリスクが 1. 25 倍になるものがあるとします。だからといって 100 人産んだら、1. 25 の 100 乗になる、というのは非現実的ですよね。病院への受診回数などもこちらになります。

Continuous variable

ある一定の範囲内の値をとるものです、その一つ一つの変化量は一定です。年齢や体重、血圧がこれにあたります。

大きくは Categorical variable か Continuous variable というに覚えることが多いのではないでしょうか。

ちょっと一コマ

サンプル数の N と n って？

また、論文などでサンプル数を記載する際、大文字の N と小文字の n を使用し、N=10,000 であったり、n=100 と記載していることがあります。一般的には母集団の数を示す場合に N を使用し、サンプル数（選び出したサンプル）を示す場合に n を使用します。

1-2 得られたデータはそのまま使用するのか？

ある研究を行い、たくさんのデータを得たとします。それらのデータをすべて使い、データをまとめることや、解析をすればいいのでしょうか？

必ずしも答えは yes とは限りません。

●エラーを見つける

これはヒトという一定ではない集団を対象とした研究で起こるもので、どのような値にも外れ値を示すことがあります。それは測定機器のエラーによるもの（偶然にも値が 10 倍を示した場合など）と、アンケートに記入してもらった際に、極端な答え方をされた場

合です。このようなケースを除外せずに解析に入れた場合、その極端なケースによって差が見えなくなることや、あるいは差が出てしまう、などのように結果に大きな影響を与えることがあります。

それぞれの集団を平均化するにはそれらの外れ値を示した対象者を解析前に除外することもあります。具体的には箱ひげ図を作成し、外れ値を除外する方法などあります。（後述）

●適切な統計解析方法を選択する

統計の本を見るとたくさんの検定法が記載されていると思います。スタンダードの検定法はよく知られていると思いますが、まずは自分の行おうとしている解析が、その検定でよいのかを初めにしっかりと考えないといけません。検定法を誤ったがゆえに差が出なかった、あるいは差が出てしまった、ということもあり得ます。後の章でも記載しますが、サンプル数が適切であったかの評価も重要です。

もっともやってはいけない方法は、データを得た後、とりあえず統計ソフトにてデータベースを作成し、たくさんの解析法を試し、差が出る検定法を探すことです。P 値が 0.05 を切った時点で有意差ありとして、それをもとに論文作成に取り掛かろうとすることです。P 値が何を意味しているかについては次の章で詳しく述べています。

その他にもたくさんの統計用語がありますが、それぞれ登場するところでも述べていきます。