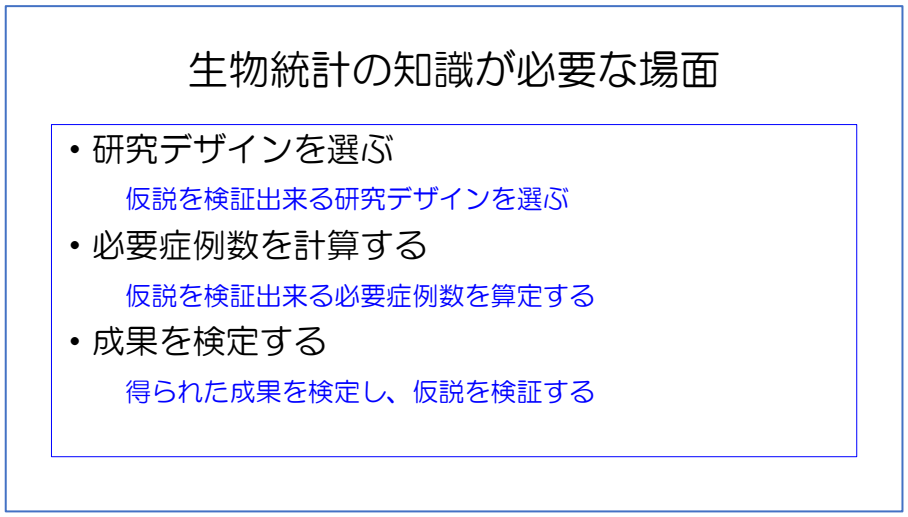
臨床研究の作り方

がん生物統計の基礎（1）

この項では JACCRO **臨床研究ワークショップ**の小講義で**竹内円雅先生・朴　慶純先生**が用いたスライドを基に「がん生物統計」の基礎的事項を解説します。詳細については別途「臨床研究の作り方」の項、あるいは「Q&A」を参照して下さい。

臨床研究において生物統計が必要な場面

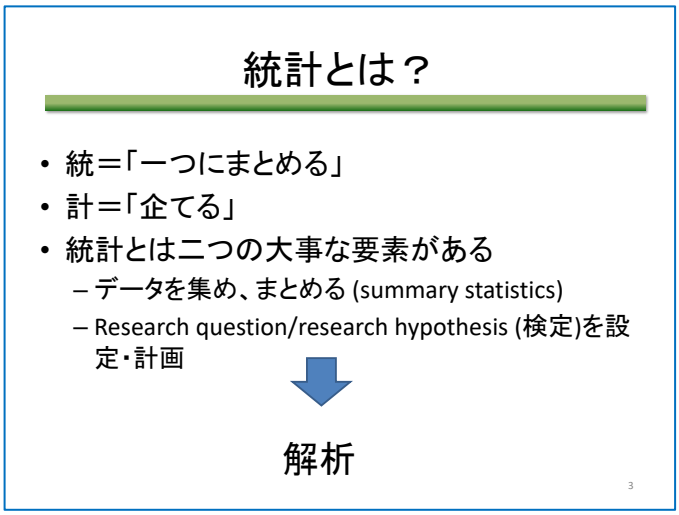


統計 Statistics は苦手と言う方が多いと思います。分からないことは統計家 Statistician に聞けば良いと考えていませんか？

左図のように臨床研究を遂行するにはあらゆる場面で生物統計の知識が必要になります。　統計家のお仕事は多岐に渡りますが、相談すべきは「仮説が検証できるか否か」の 1 点です。

また、参考となる**論文の吟味を行う場合、正しい成果であるかを自分で判断するためにも統計の知識が必要**になります。P 値、95%信頼区間、ハザード比などの意味を正しく理解する事が大事です。

統計とは



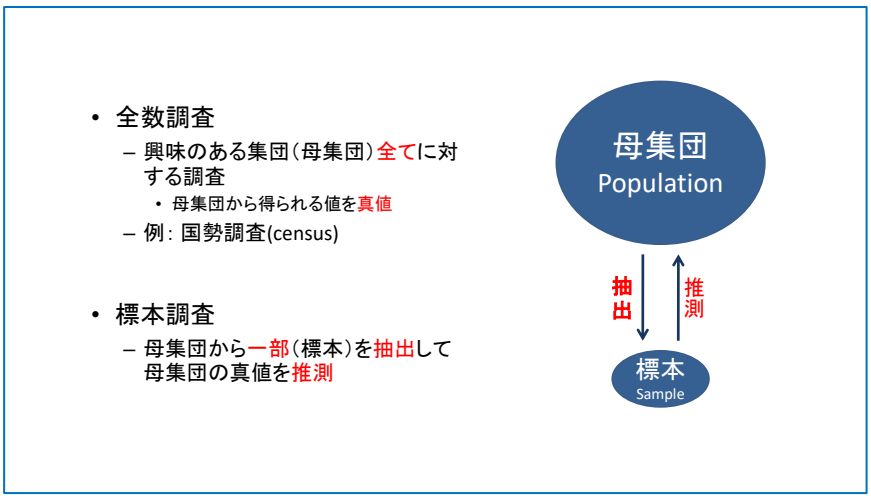
統計の「**統**」は「一つにまとめる」「**計**」は「企てる」で、全てを集めて計算すると言う意味になります。

統計には二つの大事な要素があり、その一つが「**データを集め、まとめる**」ことになります。

有名なナイチンゲールは看護師でしたが、同時に優れた数学者、統計学者でした。クリミヤ戦争の戦死者・傷病兵の膨大なデータを分析して、死因は戦闘で受けた傷では無く、受傷後の治療や病院の衛生状態の不備によるものであることを明らかにして、傷病兵の死亡率を劇的に軽減させました。現在の臨床でも過去の症例のデータベースを作って、病因や治療法の効果を検討する後向き症例集積研究などがこれに当たります。

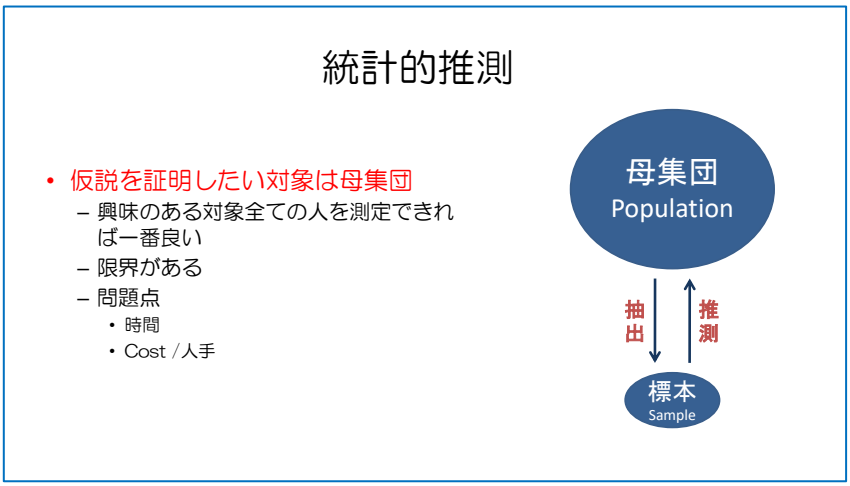
もう一つの大事な要素は「**仮説検定における解析**」です。臨床上の疑問 clinical question から「仮説 hypothesis」を導き出し、臨床研究を遂行し成果を解析して仮説が正しかったかどうかを判断する場合に統計の知識が必要になります。

なぜ統計学が必要か？

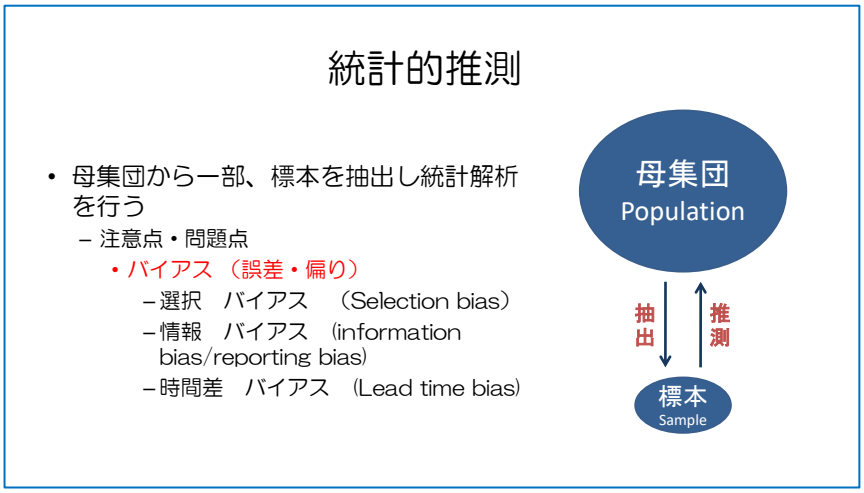


現在、ありとあらゆる領域で統計が活用されています。最も信頼性が高いのは全数調査であることは言うまでもありません。5 年に一度行われる全国民を対象とした国勢調査が代表的な例になります。対象となる集団の全例を調査するには膨大な時間と費用が掛かるので、母集団から一部を取り出して調査することを標本調査と言い、選挙の時の出口調査を含めた情報から当選確実をいち早く発表するのも標本調査の成果です。

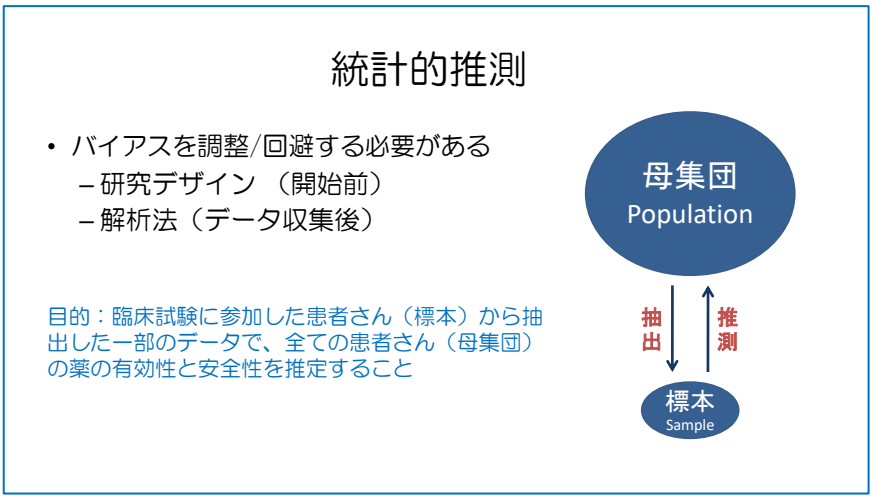
統計学的に推測



仮説を証明したい対象は母集団です。例えば胃癌 Stage III と言う集団全てで仮説を証明できれば一番良いのですが、不可能です。また出来たとしても時間、費用などが膨大になります。

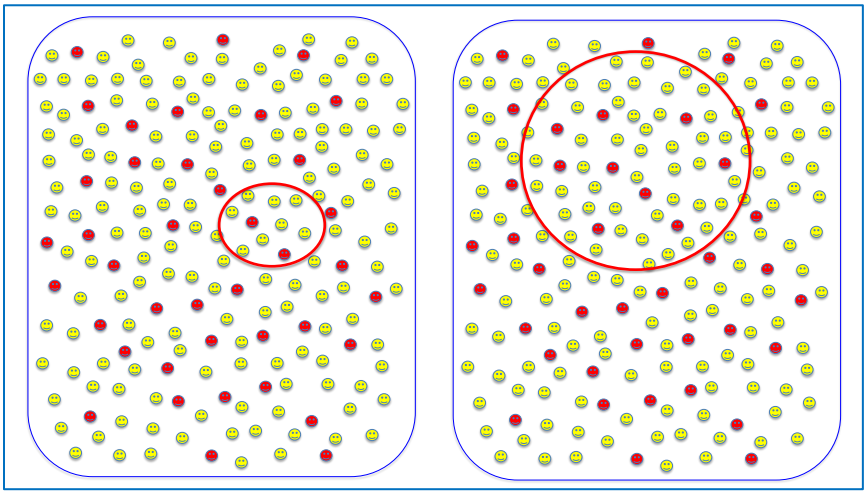


時間や費用を抑えて抽出した標本から母集団を推測する場合の注意点・問題点として「バイアス（誤差）」があります。バイアスの種類として選択バイアス、情報バイアスなどがありますが、別の項で解説します。また、このバイアスを調整・回避する様々な工夫も行われます。



バイアスを回避するために適格条件・除外条件、層別化などの工夫や、データ収集後に不適格例を除外したりします。

抽出する標本数



抽出する標本数（n）が多ければ多いほど母集団を正確に推測する事ができますが、いかに少ない n で必要十分な推測が可能なのか、95%信頼区間や必要症例数の算定などの統計学的知識が必要になって来ます。

臨床で比較するデータ

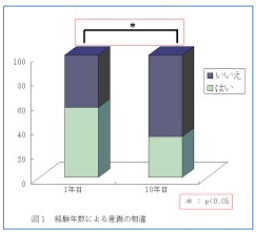
臨床で比較するのは？

比較するもの	例
率（割合）	治療の有効率 発生率（合併症など）
平均値	血圧、血糖値などの数値 出血量、輸液量、薬剤量等の数値
生存率(rate)	5年生存率 1年無病期間
順位（カテゴリー）	満足度・熱傷度・意識レベル 病期

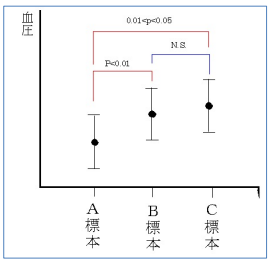
臨床研究で比較するデータには「率（割合）percent」、「平均値・中央値・最頻値」、「生存率 rate」、「順位」などがある。

連続値データ（精度の高い測定法を用いればいくらでも正確な値が得られる）、カテゴリカルデータ、イベントデータなどと分類されることもある。

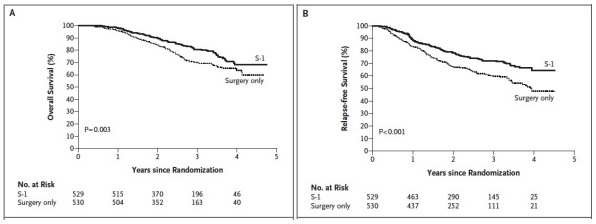
率(パーセント)の比較



平均値の比較



生存率(Rate)の比較



実際の臨床研究の論文ではこのような図で表されることが多い。

データの平均値・中央値・最頻値

平均値meanと中央値median

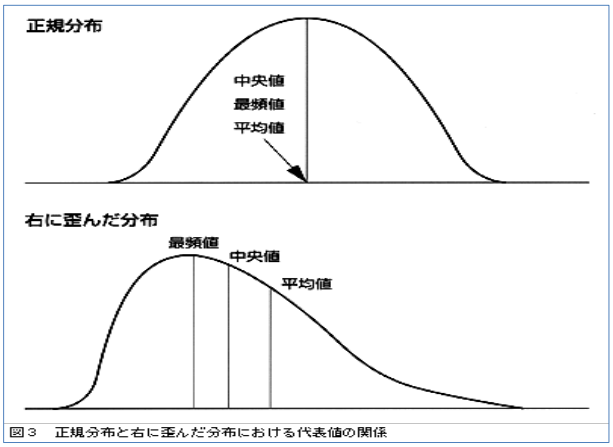


図３　正規分布と右に歪んだ分布における代表値の関係

正規分布は平均値で比較

正規分布していなければ
中央値で比較

臨床で比較するために収集するデータにはいろいろなものがあります。集めたデータをまとめてグラフにして見ると上図のように左右対象の「**正規分布**」と右に歪んだり、左に歪んだりしている「**非正規分布**」なります。例えば「中学 3 年生の身長」を 300 人計測してグラフにすると「正規分布」になり、「国民の年収」をグラフにすると「非正規分布」になるでしょう。

平均値 Mean：すべてのデータを合計して収集した個数で割った値

中央値 Median：収集したデータを大きさの順に並べて真ん中にくる値

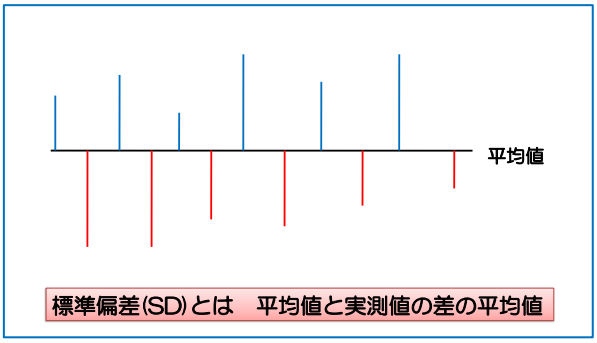
最頻値 Mode：最も集中しているデータ値（必ずあるものでは無い）

がんの臨床試験のデータの多くは非正規分布になり、中央値を比較することになります。（例：Median Survival Time: MST）

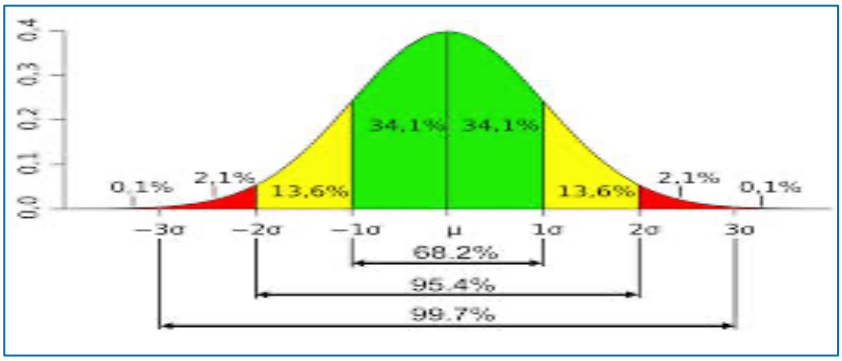
データのバラツキ：標準偏差 Standard Deviation

標準偏差とは平均値に対応するバラツキで「**平均値と実測値の差の平均値**」になります。エクセルで簡単に算出することが可能です。

SD が大きいほどバラツキが大きくなります。



平均値（中央値）± 2×SD の範囲内に全データの 95%が含まれます



がんの生物統計は難しくありません。統計学のほんの一部だけ理解できれば論文の理解、吟味が可能です。また、臨床研究の立案にも役に立ちます。がん生物統計(2)以降も続けて読んで下さい。