# 度数分布とヒストグラム

* 階級をどのように取るかを決める統一的ルールはないが、階級数に関してはスタージェスの公式が参考になる
* 所得分布のヒストグラムなどの場合、最初の階級には下限値がなく、最後の階級には上限値がない。これを「両側に開いた分布」ということがある。これらの階級については、厳密にはヒストグラムを書くことが出来ない。一般に下限値を0と見なし、上限値は適当な値を定める。  
  また、厳密な累積度数グラフを作成することもできない。便宜上、最初の階級の下限値を0にすると同時に、最後の階級については累積相対度数が1にきわみて近い点を適当に定めて点を結ぶ
* 累積相対度数のグラフについては、各階級に対して2つの異なるデータが与えられている場合（従業者別の事業所数と従業者数）、それらを組み合わせて描くことも出来る。累積相対度数を組み合わせて描いた折れ線をローレンツ曲線と呼ぶ。ローレンツ曲線は所得や資産がどの程度不平等に分配されているかを示すのに用いられる（<https://ja.wikipedia.org/wiki/%E3%83%AD%E3%83%BC%E3%83%AC%E3%83%B3%E3%83%84%E6%9B%B2%E7%B7%9A>）

# 代表値

* 幾何平均(相乗平均)(<https://www.youtube.com/watch?v=X_2Yoq5AjU0>)
  + 伸び率の平均を出すのは、算術平均よりも幾何平均の方が優れている
* 平均、メディアン、モードの関係として以下の関係が成立する。  
  一般的に、異常値に大きく左右されない点でメディアンが優れているといわれている
  + 分布が左右対称な場合 mean=median=mode
  + 左に歪んだ場合 mean<median<mode
  + 右に歪んだ場合 mean>median>mode
* 調和平均

# 相関係数

* 平均偏差も標準偏差も、ともに分布の散らばりの程度を示す指標であるが、平均偏差が用いられることはほぼない。
  + 元々の観測値が明らかにされておらず、度数分布表だけが存在する時でも、標準偏差は計算をすることが可能である。
* 分布の中心の位置が異なる場合、2つの分布の散らばりに対し、分散あるいは標準偏差でそれらを比較することは難しい。その場合、変動係数あるいは、標準得点(標準化変量)を用いることになる。

# 見かけ上の相関と編相関係数

* ある地域aにおける飲食店数と金融機関数は常識的に考えると相関関係はない(少ない)。しかし統計上、相関関係があるとみなせる場合、昼間人口を挟んで2者間に見かけ上の相関が発生している可能性を考慮しなければならない(昼間の人口が多いため、飲食店や金融機関の立地が似通る)。このようば場合には相関係数ではなく、編相関係数を用いた方がよい。
* 全体では相関がないが、それを各グループに分けた場合、その中で相関が現れることがある(イギリス全土の保守党得票率と失業率に相関関係は見られないが、イングランドのみに絞ると相関関係がみられる)。グループ分けは層別と呼ばれ、相関関係を調べる上で注意が必要となる。
* 積率相関係数は、対象データが共に量的変数である場合に用いられるが、これに対して順位相関係数は2つの順位付けの間の相関を示す指標である。スピアマン、あるいはケンドールの定義が用いられることが多い。順位が同順位となる場合は、一定の方式でこれに対する修正(タイ修正)を行う必要がある。

# 直線及び平面への当てはめ

* 回帰係数aと相関係数rの間にはa=(Sy/Sx)rの関係があり(aはrの定数倍)、xとyの相関関係は、xとyの直線関係の当てはまりの良さという意味も持つ。r^2は決定係数と呼ばれ、x(説明変数が)が(被説明変数)yを決定する強弱の度合いを表す

# 確率の定義

* ラプラスの定義では、各標本点が「同程度に確からしく」起こると仮定しており、これは証明できないが、他に反対の十分な理由がないため、起こることが妥当と考えることが出来る。これを一般に理由不十分の原則という
* ラプラスの定義を、より実際的な定義とするのが、確率の頻度説である。

# 確率変数の期待値と分散

* 期待値の性質 <https://bellcurve.jp/statistics/course/6714.html>