第 I 部

統計学

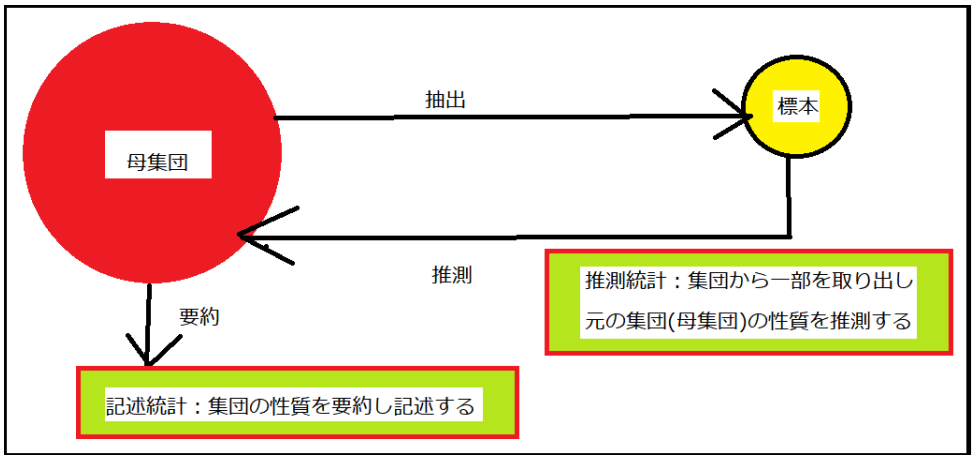
1　記述統計と推測統計

記述統計と推測統計という言葉がある

・記述統計：集団の性質を要約し記述する

・推測統計：集団から一部を取り出し元の集団 (母集団) の性質を推測する

記述統計は全データ (全傾向を含んだデータ) が揃っている場合にその性質を要約
推測統計は母集団のデータ収集が不十分な場合でも母集団の傾向を推定する様な方法を示す



2　確率変数と確率分布

確率変数とは

・事象と結びつけられた数値

・事象そのものを指すと解釈する場合も多い

確率分布とは

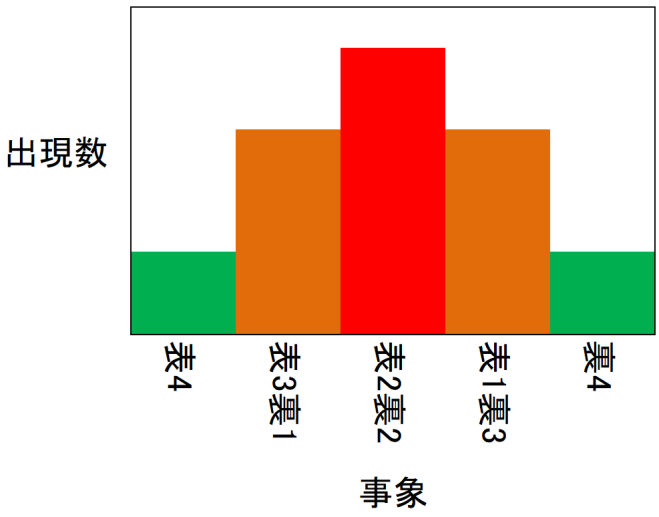
・事象の発生する確率の分布

・離散値であれば表に示せる

例えば、コインを 4 枚投げて裏表の組み合わせは 5 通りになるが、パタン (表が 4 枚、表 3 裏 1、表 2 裏 2、表 1 裏 3、裏が 4 枚) を事象といい、それらを変数化 (単に enum 化すれば良いし、表 (または裏) の出数で纏めても良い) したものが確率変数と定義されている様である

確率分布はその事象の発生する確率の分布、コイン 4 枚のトスで一番出やすいのが表 2 裏 2 で一番出難いのが表 4 または裏 4 という事でふつうは正規分布っぽい結果になるはず (コインにイカサマ仕込んでいたらその限りではない)

以下が確率分布を示したヒストグラムという事になる (実際はサンプル積まないときれいな分布にはならない)



3　期待値

期待値とは、その分布における確率変数の平均 (見えているデータの分布に偏りがある場合は平均ではなく「ありえそう」な値)

事象 X	X_1	X_2	.....	X_n
確率変数 f(X)	f(X_1)	f(X_2)	.....	f(X_n)
確率 P(X)	P(X_1)	P(X_2)	.....	P(X_n)

確率変数が「事象発生時の結果」であり、それが起こる確率がわかっている場合、期待値が求まる

(確率 P(X) で事象が発生し、それによって利益 f(X) を得るならば、期待値は単純に $f(X) * P(X)$)

離散系における期待値は

$$\text{期待値 } E(f) = \sum_{k=1}^{n} P(X_k) f(X_k)$$

連続系における期待値は

$$\text{期待値 } E(f) = \int P(X_k) f(X_k) dx$$

4　分散と共分散

分散とは

・データの散らばり具合

・データのそれぞれの値が期待値からどれだけズレているのかを平均したもの

$$\text{分散 } Var(f) = E((f(X_k) - E(f))^2) = E(f^2(X_k)) - (E(f))^2$$

共分散とは

・２つのデータ系列の傾向の違い

　・正の値を取れば似た傾向

　・負の値を取れば逆の傾向

　・ゼロを取れば関係性に乏しい

$$\text{共分散 } Cov(f, g) = E((f(X_k) - E(f))(g(Y_k) - E(g))) = E(fg) - E(f)E(g)$$

5　分散と標準偏差

分散の平方根 (分散は 2 乗なので元のデータと単位を合わせる) を標準偏差という

$$\text{標準偏差 } \sigma = \sqrt{Var(f)} = \sqrt{E((f(X_k) - E(f))^2)}$$

6　確率分布

様々な確率分布について説明

ベルヌーイ分布

・コイントスのイメージ

・裏と表で出る確率が等しくなくとも扱える

$$P(x|\ \mu) = \mu^{x}(1-\ \mu)^{1-x}$$

マルチヌーイ (カテゴリカル) 分布

・さいころを転がすイメージ

・各面の出る確率が等しくなくとも扱える

二項分布

・ベルヌーイ分布の多試行版

$$P(x|\ \lambda, n) = \frac{n!}{x!(n-x)!}\lambda^{x}(1-\ \lambda)^{1-x}$$

ガウス分布

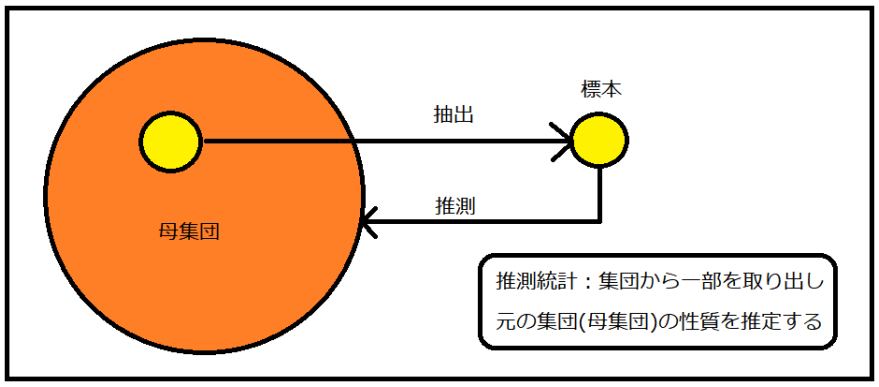
・釣鐘型の連続分布

・真の分布がわからなくてもサンプルが多ければ正規分布に近づく

$$N(x;\ \mu,\ \sigma^{2}) = \sqrt{\frac{1}{2\pi\ \sigma^{2}}}exp(-\frac{1}{2\ \sigma^{2}}(x-\ \mu)^{2})$$

7　推定

母集団を特徴づける母数 (パラメーター:平均など) を統計学的に推測すること



・点推定：平均値などを１つの値に推定すること
・区間推定：平均値などが存在する範囲 (区間) を推定すること

8　推定量と推定値

推定量 (extimator)：
パラメータを推定する為に利用する数値の計算方法や計算式のこと。推定関数ともいう。

推定値 (estimate)：
実際に試行を行った結果から計算した値

真の値を θ とすると、推定値は $\hat{\theta}$ のように示す (シータハットという)

9　標本平均

母集団から取り出した標本の平均値

・サンプル数が大きくなれば、母集団の値に近づく→一致性

・サンプル数がいくらであっても、その期待値は母集団の値と同様→不偏性

E($\hat{\theta}$)= θ である場合不偏性と一致性を兼ね備えた良い標本であるといえる

10　標本分散

サンプルサイズを n として

$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})$
(ただし、$\bar{x}$ は x の平均)

自明ではあるが上記は一致性は満たすが不偏性は満たさない (データのばらつきに影響される)

11　不偏分散

不偏分散 (標本分散を修正する)

$s^2 = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})$

結局サンプル数を 1 減らしている様なトリックだけど、これは x の平均でデータの偏差を縛っているという理屈より分母をいじって分散の値を少し増やすためという事らしい (なんだかよくわからない理屈ではある、講師もそのような意味のことを言っている)

さらに強調しているのが 1 つのサンプルの重みは全サンプル数との比率で決まるという事

数式では $\frac{\Delta W}{W}$ が重み (W がサンプル数で Δ W が 1 つのサンプルを表す、これは W の積分である) と説明している、これは事象数に対する 1 つのイベントの確率であるという説明でもある