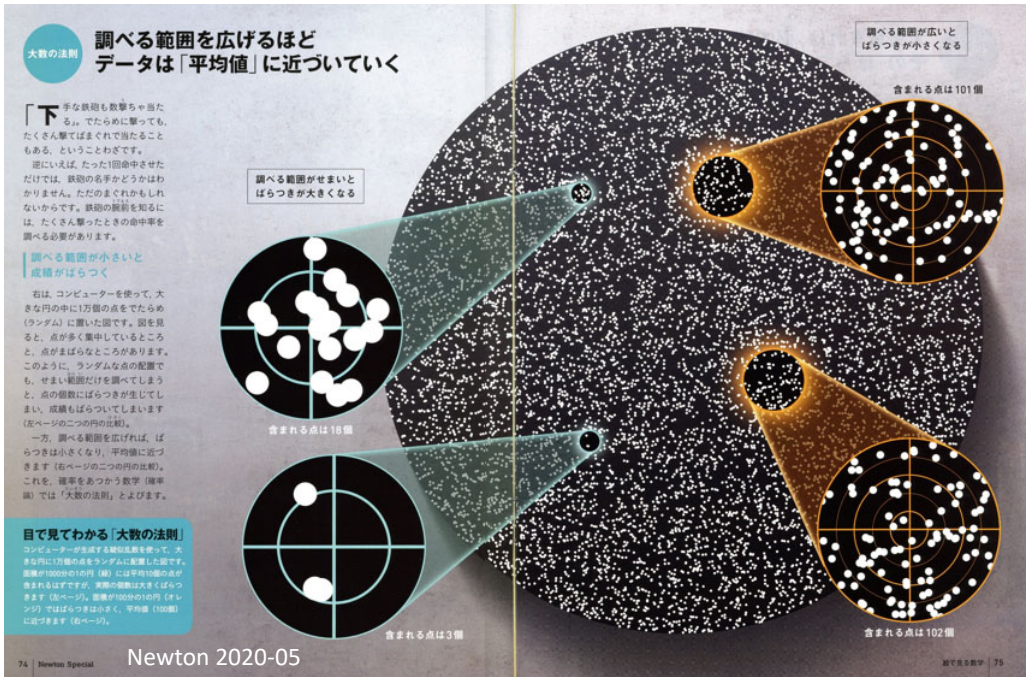
2022年度 Data Science実践B 講義資料

- ✓ 大数の法則
- ✓ 中心極限定理
- ✓ 推定
- ✓ 仮説と検定

母集団と標本 ～ 推測統計のコツ



Newton 2020-05

確率変数の振る舞い

サイコロを投げて、ある目が出る確率は1/6。
10回程度であれば、1が8回出るかもしれないし、あるいは1回も出ないかもしれない（その他の目も同様）。

➡ 試行回数（サイコロを投げる回数）が少ない場合、**目の出る回数に極端な偏りが出る。**

➡ 試行回数を1000回、10万回と増やして行けば、ある目（どの目でも）の出る確率は1/6に近づいていく。

（客観確率 / 頻度確率）

大数の法則

ある試行を何回も繰り返していくと、出現確率は一定値に近づいていく➔ 標本平均も一定に近づく

◯ $\bar{X} \to \mu$
× $E(\bar{X}) \to \mu$

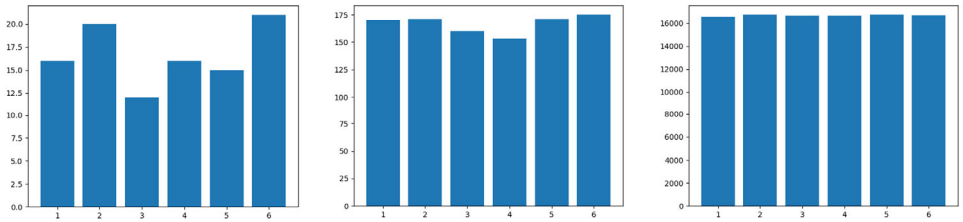
大数の法則 (law of large numbers)

標本数 n を無限大とする極限で、標本平均 $\bar{X}$ は母平均 μ に収束する

サイコロでシミュレーション

サイコロを（理想的には）無限回投げた結果、１から６の目が出る確率はそれぞれ 1/6 である。

（客観確率 / 頻度確率）



100回　1,000回　100,000回

それはそうなのだが、現実には（特に実験的観点からは）非常に使いにくい。

∵ 無限回サイコロは投げられない。（少なくとも十分な大きさのサンプル（標本＝実験の場合は同一条件下での測定）が必要

確率変数の和の期待値と分散

独立な確率変数の和の期待値および分散は、

$V(nX_i)$ との違いを理解せよ

$$E\left(X_1+X_2+\cdots+X_n\right)=n\mu$$

$$V\left(X_1+X_2+\cdots+X_n\right)=n\sigma^2$$ で与えられる。

（標準偏差は$\sqrt{n}$倍）

10枚の薄い板を重ねた合板（ベニヤ板）の厚さについて、

期待値　1枚の場合の10倍

標準偏差　1枚の場合の$\sqrt{10}$倍　となる。

独立な確率変数であれば、$E(X_i)=\mu_i,\ V(X_i)=\sigma_i^2,\ i=1,2,\cdots,n$ のとき、

$$E\left(X_1+X_2+\cdots+X_n\right)=E(X_1)+E(X_2)+\cdots+E(X_n)$$

$$V\left(X_1+X_2+\cdots+X_n\right)=V(X_1)+V(X_2)+\cdots+V(X_n)$$ が従う。

また、個々の確率変数が正規分布に従い独立である場合は、和についてもまた、正規分布 $N(n\mu,n\sigma^2)$に従う（正規分布の再生性）

確率変数の平均の期待値と分散

$X_1,X_2,\cdots,X_n$は互いに独立で、同じ平均 μ と分散 σ^2 を持つとする（分布は任意）。

このとき、 $\bar{X}=\dfrac{X_1+X_2+\cdots+X_n}{n}$

を**標本平均**とよび、やはり確率変数となる。

標本平均の期待値は、$E(X_k)=\mu\ (1\le k\le n)$ より、

$$E(\bar{X})=\frac{n\mu}{n}=\mu$$ である。

nに依らない

$n\to\infty$である必要はなく、常に成立！！

一方、標本平均の分散は、

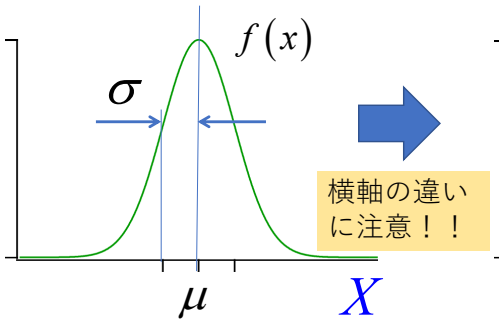
$$V(\bar{X})=V\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)=\frac{n\sigma^2}{n^2}=\frac{\sigma^2}{n}$$ となる。

$$V(\bar{X})=E\left[\left(\bar{X}-E(\bar{X})\right)^2\right]=E\left[\left(\bar{X}-\mu\right)^2\right]$$

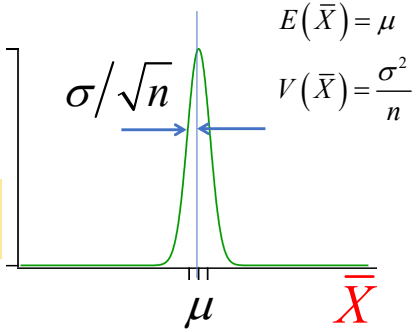
$V[aX+b]=a^2V[X],\ a=1/n$

従って平均板厚の標準偏差は、1枚の場合の**$1/\sqrt{n}$ 倍**となる。

確率変数の平均の期待値と分散



横軸の違いに注意！！



1つ1つの確率変数が従う分布（上の図は正規分布で描いているが、分布は仮定しない）

母平均：μ
母分散：σ^2

複数の確率変数さの標本平均

$$\bar{X}=\frac{X_1+X_2+\cdots+X_n}{n}$$

が従う分布

標本平均の平均：μ
標本平均の分散：σ^2/n

大数の法則の主張

前述のとおり、標本平均 $\bar{X}$ の平均が μ であること、すなわち

$E(\bar{X})=\mu$ が大数の法則**ではない**ことに注意。

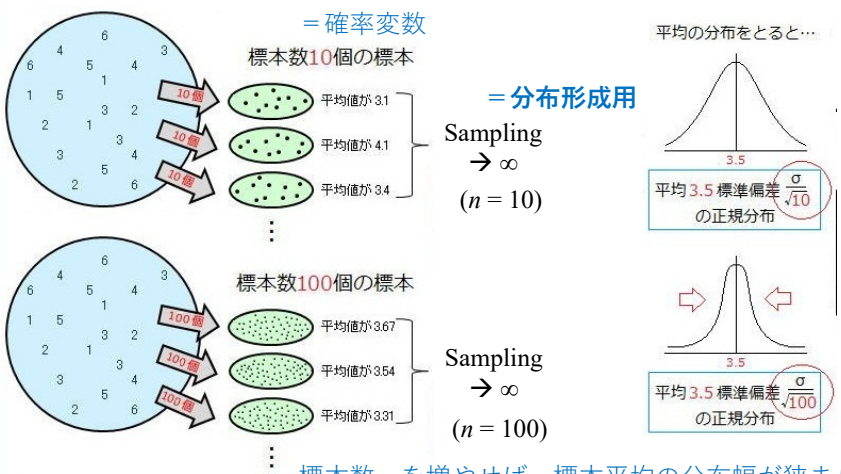
$E(\bar{X})=\mu$ は、$\bar{X}$ の値が確率に従って分布し、その分布の**期待値**が**母平均 μ に一致**することを示している。

一方、大数の法則は、$n\to\infty$ の極限では $\bar{X}$の**値そのもの**（期待値ではない！）が μ になる確率が1、つまり**必ず母平均 μ に一致**する（数学的に色々と難しい問題はあるが、今それは気にしないことにする）ことを主張している。

気になる人は、チェビシェフの不等式、確率収束等を調べてみよ。

実用上は、母平均・母分散が存在さえすれば（分布は不問）、測定回数を増やせば、標本平均は母平均 μ にいくらでも近づく、ということ。

大数の法則：まとめ

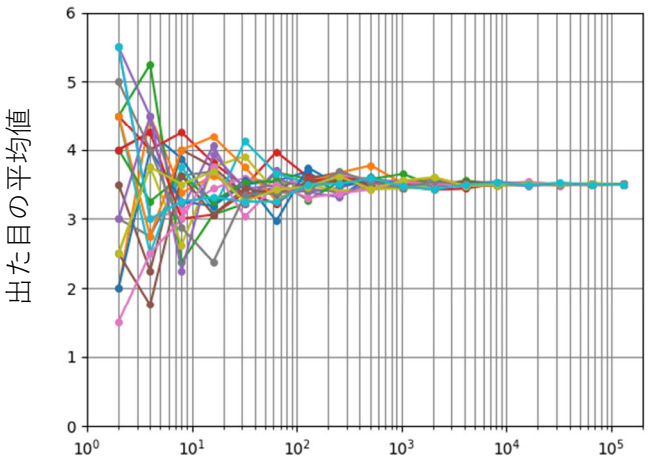


標本数 n を増やせば、標本平均の分布幅が狭まり、
標本平均が母平均 μ にいくらでも近づく。

この事実を用いれば、1回の n の大きな実験を行った結果＝母平均としてよい。

大数の法則：シミュレーション

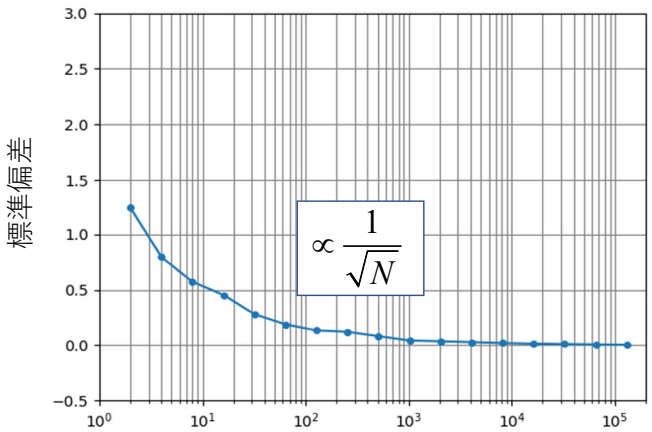
母集団から無作為抽出された**標本平均**は、標本の数をどんどん大きくしていくと、真の平均に限りなく近づいていく。



サイコロを振った回数

大数の法則：シミュレーション

母集団から無作為抽出された標本平均は、標本の数をどんどん大きくしていくと、真の平均に限りなく近づいていく。
それに伴い、**標本分散**は限りなく0に近づく。



サイコロを振った回数

大数の法則：標本平均の分散

n 回の平均を得る試行を1回とし、これを何度も繰り返して得られる標本平均

$$\bar{X} = \frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right)$$ の標準偏差

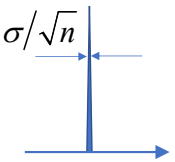
$\sqrt{V(\bar{X})}$ は、個々の標準偏差 $\sqrt{V(X_k)}$ の $\frac{1}{\sqrt{n}}$ になる。

➡ $\lim_{n\to\infty}\frac{1}{\sqrt{n}} = 0$ より、$n \to \infty$ の極限で平均周りの揺らぎは消失。

例： 1mol余り（1×10^{24}）の物資が示す統計量についての観測値の平均値周りの揺らぎは、

$$\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{1\times10^{24}}} = 1\times10^{-12} \quad \sim \sigma/\sqrt{n}$$

観測値 $= \bar{X}\left(1+10^{-12}\right)$



平均値からのズレは、1に対し0.000000000001のオーダー
→ 誤差は事実上、無視できる

物性物理、化学等の物質科学が堅牢な理由！

（参考）大数の法則の応用例

保険は大数の法則を基に成り立っている？

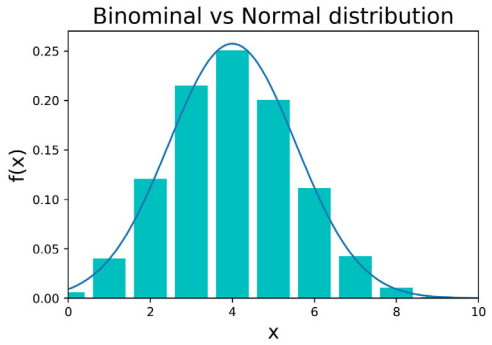
たとえば、死亡率が1,000分の1の集団があったとする。
この集団の人数が仮に1,000人だったとすると、死亡する人数の期待値は1人。
が、実際には、2人死亡するかも知れないし、ゼロかもしれない。
前者の場合、死亡した人の割合は死亡率の倍の1,000分の2で、後者の場合はゼロ。死亡者が1人のときだけ、割合が確率に一致。

この集団が100万人になると死亡者の期待値は1,000人。
ここまで人数が増えると死亡者が増えても減っても、その差はせいぜい数十人くらいで、余程のことがない限り実際の死亡者が倍の2,000人になったり、ゼロだったりすることはないだろう。つまり人数が増えることで、死亡する割合が死亡率である1,000分の1に近づいていく。

逆に死亡率が分からない時でも、この大数の法則を利用すれば、死亡率を推定できる。これが、保険は大数の法則を基に成り立っていると言われる所以。

http://media.lifenet-seimei.co.jp/2015/03/24/2761/　一部改変

正規分布



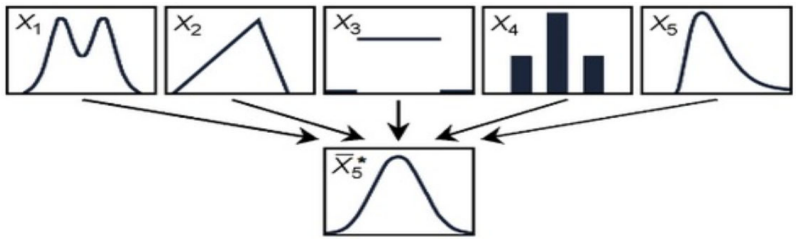
正規分布は、

$$B[n,p] = {}_nC_x p^x (1-p)^{n-x} \xrightarrow{n\to\infty} f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

により、二項分布の極限として定義されていた。

➡ ということは、元々の分布が二項分布でないと使えない？

中心極限定理



この定理は元の母集団の分布型に関する仮定を必要としない。元の分布が連続型であろうと離散型であろうと、平均と分散さえ存在すれば成立する、一般性が高く、**統計学を支える重要な定理**。

正規近似の精度は、標本数と母集団分布の形に関連する。
母集団分布が連続かつ単純な形であれば、小さい標本数でもよい近似が得られるが、母集団の分布が離散的であったり、複雑な形状の場合、より大きな標本数が必要。
多くの場合、標本数 = 20~30 が正規近似の目安と考えられる。
（卒業研究等で使う場合は、大標本として解析した、などと添えておく）

中心極限定理

中心極限定理

確率変数 $x_1, x_2, \cdots, x_n$ はそれぞれ独立で、母平均 μ、母分散 σ^2（有限値）を持つ任意の分布に従うとする。

n が十分大きいとき、その平均

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

の分布は、

平均の「値」ではない！

正規分布 $N\left(\mu, \sigma^2/n\right)$ に限りなく近づく。

（同じことだが）独立な確率変数の和バージョン

元の分布の平均を μ、分散を σ^2 とすると、確率変数の和は元の分布がなんであろうと、n が大きくなるにつれて平均 $n\mu$、分散 $n\sigma^2$ の正規分布に限りなく近づく。

中心極限定理（一般化）

各確率変数が由来する母集団の母数が異なっている場合に一般化できる。

一般化された中心極限定理

確率変数 $x_1, x_2, \cdots, x_n$ はそれぞれ独立で、母平均 $\mu_1, \mu_2, \cdots, \mu_n$
母分散 $\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2$（有限値）を持つ任意の分布に従うとする。

n が十分大きいとき、その平均

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$ の分布は、

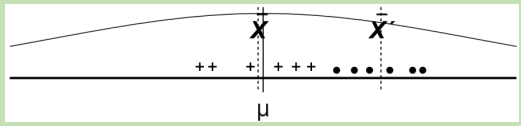
正規分布 $N\left(\frac{1}{n}\sum_{i=1}^{n}\mu_i, \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2; \bar{x}\right)$ に限りなく近づく。

重要な例外は確率関数の少なくとも1つの従う分布の分散が無限大（ローレンツ関数）の場合で、標本平均の分布は一般に正規分布にはならない。

- ✓ 大数の法則
- ✓ 中心極限定理
- ✓ 推定
- ✓ 仮説と検定

母分散と標本分散

標本と標本平均の距離の和は、標本と母平均との距離の和より小

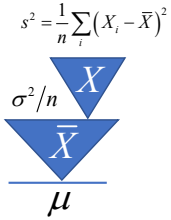


標本平均を用いて計算される標本分散は、真値（＝母分散）よりも小さく見積もられる。

標本分散の期待値は、母分散より標本平均の分散 σ^2/n の分だけ小さい。

自由度 $n-1$ について、独立変数の和の分散の性質を用いて以下のように（イメージが沸きやすい方法で）導出できる。

標本分散から母分散を推定する際、母平均が未知なので、（母平均との関連は不明だが）標本平均で代用。
X は標本平均 $\bar{X}$ の周りに標本分散を持って分布し、また標本平均 $\bar{X}$ は母平均 μ を中心に σ^2/n の分散で分布している。

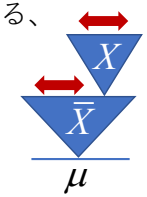


不偏分散の導出

全分散は（独立であれば）部分的な分散の和に分解できる、すなわち

$$V(X_1 + X_2) = V(X_1) + V(X_2)$$

が成り立つことを用いると（p.53の式(3.14)）、



我々が知りたい母分散は、$\bar{X}$を中心とするXの標本分散と、母平均 μ を中心とする$\bar{X}$の分散の和で書ける。

$$\sigma^2 = E\left[s^2\right] + \frac{\sigma^2}{n} \rightarrow E\left[s^2\right] = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2 \rightarrow \sigma^2 = \frac{n}{n-1}E\left[s^2\right]$$

標本分散については、

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

この両辺の期待値を取り、上の2式より、 $E\left[\sum_{i=1}^{n}\frac{(x_i - \bar{x})^2}{n-1}\right] = \sigma^2$

期待値が母数に一致するような推定量を、**不偏推定量**という ⇨ $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ **不偏分散**

不偏分散の導出（別法）

$$\sum_i (x_i-\bar{x})^2 = \sum_i \{(x_i-\mu)+(\mu-\bar{x})\}^2$$
$$= \sum_i (x_i-\mu)^2 + 2\sum_i (x_i-\mu)(\mu-\bar{x}) + \sum_i (\mu-\bar{x})^2$$
$$= \sum_i (x_i-\mu)^2 - 2(\bar{x}-\mu)\sum_i (x_i-\mu) + n(\bar{x}-\mu)^2$$
$$= \sum_i (x_i-\mu)^2 - 2(\bar{x}-\mu)(n\bar{x}-n\mu) + n(\bar{x}-\mu)^2$$
$$= \sum_i (x_i-\mu)^2 - n(\bar{x}-\mu)^2$$

各項の期待値を取ると、右辺の２項は

$$E\left[\sum_i (x_i-\mu)^2\right] = \sum_i E\left[(x_i-\mu)^2\right] = n\sigma^2$$
$$E\left[n(\bar{x}-\mu)^2\right] = nE\left[(\bar{x}-\mu)^2\right] = nV[\bar{x}] = n\frac{\sigma^2}{n} = \sigma^2$$

不偏分散の平均＝母分散

$$\therefore E\left[\sum_i (x_i-\bar{x})^2\right] = n\sigma^2 - \sigma^2 = (n-1)\sigma^2 \Rightarrow E\left[\frac{1}{n-1}\sum_i (x_i-\bar{x})^2\right] = \sigma^2$$

母分散と標本分散

$$\sigma^2 = \frac{1}{n}\left\{(X_1-\mu)^2 + (X_2-\mu)^2 + \cdots + (X_n-\mu)^2\right\}$$

同じ μ、σ² から生まれる標本は無数にある、ということ

母平均 μ と母分散 σ² は、分布を指定する独立な（ある定まった）数値。分布から生み出される具体的な標本とは独立。

$$s^2 = \frac{1}{n}\left\{(X_1-\bar{X})^2 + (X_2-\bar{X})^2 + \cdots + (X_n-\bar{X})^2\right\}$$

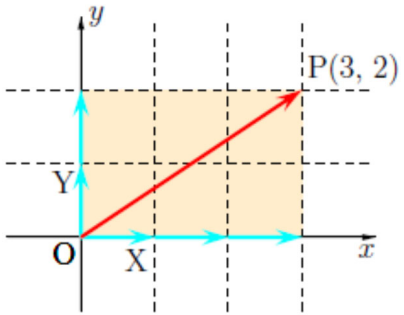
標本平均や標本分散は、標本（データ）の採り方により変わる数値であり、標本平均を決めるとデータは自由に選べない → 自由度1減

n–1の意味は、統計量のもつ以下の基本的性質とも関係する。
母平均は母集団分布の位置を表す母数であり、極端だが1つの標本で足りる。
例えば大人と小学生の体重を比べる場合、最悪一つの測定値で推測しうる。
一方、ある小学生の体重が20 kg、ある大人の体重が60 kgであった。小学生と大人の集団でどちらのバラツキが大きいか、については1点では全く情報が得られない。少なくとも2つの測定点が必要である。
同様に、歪み（非対称性）は3点ないと決まらないし、とがり具合（尖度）を推測しようとすれば、最低4点必要となる。

自由度について

自由度：独立に指定しうる変数の数

例：2次元平面上で1点を指定する場合



例えば Y = X

- 何の制約もない場合：2
- XとYの間に、 Y=f(X)なる何らかの関係がある場合：1

自由度について

$X_1, X_2, \cdots, X_n$ ：独立

⇩

$X_1-\mu, X_2-\mu, \cdots, X_n-\mu$ も独立

$\bar{X} = \frac{1}{n}(X_1+X_2+\cdots+X_n)$ が既知の場合、これが確率変数間の制約条件となる。

具体的には、この標本平均との偏差

$X_1-\bar{X}, X_2-\bar{X}, \cdots, X_n-\bar{X}$ をつくると、これらの和

$$(X_1-\bar{X})+(X_2-\bar{X})+\cdots+(X_n-\bar{X})$$
$$=(X_1+X_2+\cdots+X_n)-n\bar{X}$$
$$=0$$

より、独立ではない。

未知量 μ を使っても独立を保つが、既知量 $\bar{X}$ を使うと自由度が1減る

第７章　推定　例題（母平均の推定）

ある種のハエを20匹無作為に抽出して観察を続けたところ、寿命の平均が28日、標準偏差は4日であった。このとき、この種のハエ全体の平均寿命（母平均）μについて、90%信頼区間と99%信頼区間を推定せよ。

また、標本としてハエ100匹を抽出した場合についても推定してみよ。

仮説と検定

「カラスが黒い」
ことを証明せよ

仮説・検定のココロ

命題：正しい（真）か正しくない（偽）かが数学的・論理学的に定まる式や文章。

帰無仮説：

主張したい命題(A)の否定$(\overline{A})$を仮定する。
$\overline{A}$は（願わくば）否定され、無に帰すことになる。
$\overline{A}$を否定することで、元々主張したかった命題Aに寄せていく。

なに故こんなややこしいことをするのか？

直接的に命題が真であると示すのは容易ではない。
無限回試行して、その全てが仮説を満たしてはじめて真となる。
有限回の試行で仮説を満たしたとしても、次の試行で満たさないかもしれず、強い結論は得られない。
一方、一度でも反例が出てくればその命題は偽。（強い結論）

仮説・検定のココロ

命題：「カラスは黒い」（カラスは黒 or 白、を仮定）

ある瞬間に存在する地球上のカラスを漏れなく調査し、全てが黒であれば真となる。

非現実的

一方、少数の調査であっても白カラスが1羽でも見つかれば、その時点で命題「カラスは黒い」は偽となる。

たとえ100億羽調べて全て黒としても、100億1羽目が白かもしれない。この場合、「カラスは黒い」は、真とは言えない、なる曖昧で弱い結論（？）しか得られない。

発想を変えてみる

主張したいことの否定命題：「カラスは白い」は、黒いカラスを1羽連れてくれば否定できる。（強い結論）
否定命題を棄却することで、「カラスは黒い」に寄せる。
（この場合でも、「カラスは黒い」とは言えない。。。）