☆生物実験データの統計学

ここでの狙い：

・**「ゆらぎ」を含んだ量をいかにして扱うかの感覚を掴む**

＊実験データの誤差

＊細胞内化学反応ダイナミクスのゆらぎ

・「ゆらぎ」を含んだ量を、単に代表値（例えば平均値）のみで記述するのではなく、**その背後にある分布を通じて理解する**

＊正規分布の重要性（中心極限定理）

・さらには「ゆらぎ」を含んだ2つの量を比較する方法を学ぶ

＊t検定、χ^2検定

☆生物実験データの統計学

例えば、ある人があるたんぱく質の濃度を同じように3回測定したが、結果が以下のようになった。

1回目：　0．12　mol／ℓ
2回目：　0．19　mol／ℓ
3回目：　0．08　mol／ℓ

その人はこの3回の結果を平均して、0．13mol／ℓという値を得た。
しかし、別にまた3回測定したら、別の平均値になりそう、、、

もっと測定回数を増やして平均値を求めれば良いような気もするが、それは何故？？　どれくらい回数を増やせばどれくらい良くなる？

☆コイン投げで平均値の性質を調べる

一回（あるいは少ない回数）の実験では結果がゆらいでしまって、
正しい結果は得られそうもない。

⇧

では、数多くの実験を行ってその平均値をとれば、
正しい結果が（直感的には）得られそうだが、それは本当？

この問いを考えるため、以降ではコイン投げのプロセスを例として
平均値の性質を議論する。

偏りのないコインを投げて、表が出れば1、裏が出れば0の結果が得
られるものとする。当然、コインを一回投げるという実験の期待値は

$$1\times\frac{1}{2}+0\times\frac{1}{2}=\frac{1}{2}$$

となる。一方で、この実験を繰り返して平均値をとれば、この正しい期
待値になると期待されるが、何回くらい実験を行えば、どれくらいの精
度でこの期待値が得られるであろうか？

☆コイン投げの統計

問い：

確率1／2で表が出る（当然、確率1／2で裏が出る）コインをn回投げて、
表が出た数を記録する。このとき、表がk回出る確率は？

考え方：

コインをn回投げて、表がk回となる可能な組み合わせの数を数え上げて、
その場合の数にそれが実現される確率の積を求める。

回答：

表がk回出る確率をP(k)とすると、

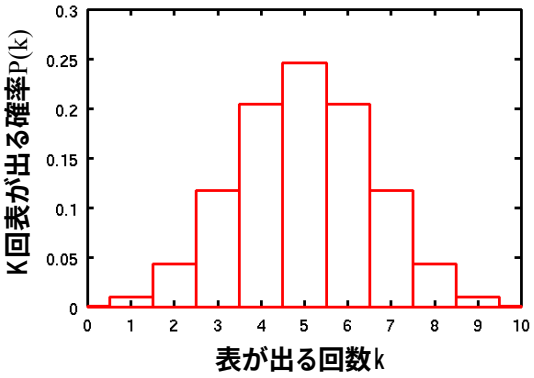
$$P(k)={}_nC_k\left(\frac{1}{2}\right)^k\left(\frac{1}{2}\right)^{n-k}$$

ただし、 ${}_nC_k=\dfrac{n!}{k!(n-k!)}$

となる。このような分布を二項分布（またはベルヌーイ分布と言う）。

☆n=10の場合の二項分布

コインを投げる回数nを１０とすると、k回表が出る確率P(k)
は以下のようになる。



このグラフから、投げる回数nが１０の場合は、実験結果の平均値をとることによって、正しい期待値（k／n＝０．５）が得られる確率は約０．２５であり、あまり大きくないことが判る。

☆二項分布の期待値と分散（と標準偏差）①

コインを投げたときに表が出る確率をpとし、裏がでる確率をqとする。
（つまり、p＋q＝1） このとき、n回コインを投げて表が出る回数kの
期待値E（k）と分散V（k）は以下のように計算できる。

$$E(k)=\sum_{k=1}^{n}kP(k)=\sum_{k=1}^{n}k\,{}_nC_k\,p^k q^{n-k}$$

ここで、以下の二項定理を利用する。

$$(p+q)^n=\sum_{k=1}^{n}{}_nC_k\,p^k q^{n-k}$$

この両辺をpについて微分し、さらに両辺にpを掛けると、以下のようになる。

$$n(p+q)^{n-1}p=\sum_{k=1}^{n}{}_nC_k\;kp^{k-1}q^{n-k}p=\sum_{k=1}^{n}{}_nC_k\;kp^k q^{n-k}=E(k)$$

ゆえに$p+q=1$より、 $\boxed{E(k)=np}$

☆二項分布の期待値と分散（と標準偏差）②

次に、二項分布の分散を計算してみる。分散V(k)とは、表が出る回数kとその期待値の差 (k－E(k)) の2乗の期待値と定義される。E(k)=μ とすると、

$$V(k)=E\left[(k-\mu)^2\right]=E\left[k^2-2k\mu+\mu^2\right]$$
$$=E(k^2)-2\mu E(k)+\mu^2=\underline{E(k^2)-\mu^2}$$

つまり、分散とはkの2乗の期待値からkの期待値の2乗を引いたものとなる。ここで、kの2乗の期待値とは、以下のように表される。

$$E(k^2)=\sum_{k=1}^{n}k^2P(k)=\sum_{k=1}^{n}k^2\,{}_nC_k\,p^kq^{n-k}$$

これは、先程と同様に二項定理 $(p+q)^n=\cdots$ の両辺をpについて（こんどは）2回微分し、さらに両辺に p^2 を掛けることによって計算できる。

☆二項分布の期待値と分散（と標準偏差）③

$$n(n-1)(p+q)^{n-2}p^2=\sum_{k=1}^{n}{}_nC_k\,k(k-1)p^{k-2}q^{n-k}p^2$$

$$n(n-1)p^2=\underbrace{\sum_{k=1}^{n}k^2\,{}_nC_k\,p^kq^{n-k}}_{E(k^2)}-\underbrace{\sum_{k=1}^{n}k\,{}_nC_k\,p^kq^{n-k}}_{E(k)=np}$$

これをまとめると、

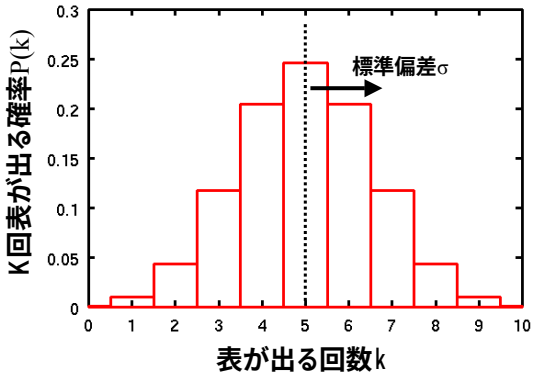
$$E(k^2)=n(n-1)p^2+np$$

これらから二項分布の分散V(k)と標準偏差σは、

$$V(k)=E(k^2)-\left(E(k)\right)^2=n(n-1)p^2+np-n^2p^2$$
$$=np(1-p)=\underline{npq}$$
$$\sigma=\sqrt{V(x)}=\sqrt{\underline{npq}}$$

☆標準偏差とは何だったか

二項分布の標準偏差は $\sqrt{npq}$ となる。このことは、分布の期待値からの**ずれの平均**がその程度であることを意味する。



右の図の場合、n=10, p=1/2, q=1/2なので、標準偏差σは

$$\sqrt{10 \cdot \tfrac{1}{2} \cdot \tfrac{1}{2}} = \sqrt{2.5} \approx 1.58$$

これまでの議論から、コイン投げの表の出る期待値は回数nに比例し、その分布の幅（≈ 標準偏差）はnの平方根に比例することが判った。
では、それは何を意味するであろうか？

☆二項分布はnが大きいときに正規分布で近似できる

二項分布は正確に計算できるが、n（この場合は、コインを投げる回数）が大きくなると計算が大変になり、いろいろな意味で便利でない。そこで、nが大きいときの便利な近似として、以下のものを使う。

期待値 $\mu = np$, 標準偏差 $\sigma = \sqrt{npq}$ を持つ以下の二項分布

$$P(k) = {}_nC_k p^k q^{n-k}$$

は、nが大きいときに以下の**正規分布**

$$P(k) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(k-\mu)^2}{2\sigma^2}}$$

に従う。（当然、この正規分布も期待値 $\mu = np$ 、標準偏差 $\sigma = \sqrt{npq}$ を持つ）

☆二項分布→正規分布の証明（ちょっと大変）①

この証明では、以下の2つの関係式を使う

- **スターリングの公式**

$$\log n! \approx n\log n - n$$

- **テイラー展開**

$$f(a+x) = f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + \cdots$$

これらの関係式を用いて、二項分布を展開していく。まず、話を簡単にするために、二項分布の対数を取って、それを変形していく。

$$P(k) = \frac{n!}{k!(n-k!)}p^k q^{n-k}$$

$$\begin{aligned}\log P(k) &= \log n! - \log k! - \log(n-k)! + k\log p + (n-k)\log q \\ &= n\log n - n - k\log k + k - (n-k)\log(n-k) + (n-k) \\ &\quad + k\log p + (n-k)\log q\end{aligned}$$

☆二項分布→正規分布の証明（ちょっと大変）②

$$\log P(k) = n\log n - k\log k - (n-k)\log(n-k) + k\log p + (n-k)\log q$$

ここで、 $\log P(k)$ **を** k=np **の周りでテイラー展開するために、**$\log P(k)$ **をkで微分、二階微分したものを計算する。**

$$\frac{d\log P(k)}{dk} = \log\left(\frac{n-k}{k}\right) + \log\left(\frac{p}{q}\right)$$

$$\frac{d^2\log P(k)}{dk^2} = -\left(\frac{1}{n-k} + \frac{1}{k}\right)$$

上の微分係数にk=npを代入すると

$$\left.\frac{d\log P(k)}{dk}\right|_{k=np} = 0 \quad , \quad \left.\frac{d^2\log P(k)}{dk^2}\right|_{k=np} = -\frac{1}{npq}$$

☆二項分布→正規分布の証明（ちょっと大変）③

$$\log P(k) = n\log n - k\log k - (n-k)\log(n-k) + k\log p + (n-k)\log q$$

ここで、$\log P(k)$ を k=np の周りでテイラー展開する。

$$\log P(k) = \log P(np) + \left.\frac{d\log P(k)}{dk}\right|_{k=np}(k-np) + \frac{1}{2}\left.\frac{d^2\log P(k)}{dk^2}\right|_{k=np}(k-np)^2 + \cdots$$

$$= \log P(np) - \frac{1}{2}\cdot\frac{1}{npq}(k-np)^2 + \cdots$$

ここで、3次以降の項は、$n^{-\alpha+1}$ （α は次数）のファクター
が入ってくるのでnが大きいときには無視できる。

☆二項分布→正規分布の証明（ちょっと大変）④

$$\log P(k) \approx \log P(np) - \frac{1}{2}\cdot\frac{1}{npq}(k-np)^2$$

$$P(k) = Ae^{-\frac{(k-\mu)^2}{2\sigma^2}},\ \mu = np,\ \sigma = \sqrt{npq},\ A \text{ は定数}$$

Aは、P(k)を全ての可能な k にたいして和を取ると1になる（確率であるために）
という条件から、

$$\int_{-\infty}^{\infty} P(k)dk = \int_{-\infty}^{\infty} Ae^{-\frac{(k-\mu)^2}{2\sigma^2}}dk = 1,\quad \int_{-\infty}^{\infty} e^{-x^2}dx = \sqrt{\pi}$$

これを解いてまとめると、

$$P(k) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(k-\mu)^2}{2\sigma^2}}$$

ガウス積分と呼ばれる
有名な公式！

☆二項分布は正規分布によってよく近似できる

下の図に見られるように、nが10程度より大きくなると、二項分布と正規分布にはほとんど差はない。

deleted based on copyright concern.
unspecified quotation.

n=10 の場合　　n=30 の場合

二項分布
正規分布

☆正規分布の基本的な性質

期待値μ、標準偏差σの正規分布

$$P(k)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(k-\mu)^2}{2\sigma^2}}$$

においては、μ±σに全体の約68%のデータが入る。

また、μ±2σには全体の約95%のデータが入る。

deleted based on copyright concern.
unspecified quotation.

☆コイン投げの平均値の分布はどうなる？

これまでの考察により、n回コインを投げたときにk回表が出る確率は

- 二項分布（nが大きい時には近似的に正規分布）に従い、
- その期待値は np
- 標準偏差は $\sqrt{npq}$

となることが見出された。

つまり、nを増加させていくと、**期待値はnに比例して増加**するが、
標準偏差（分布の幅）はnの平方根に比例して増加する。

では、（最初の問題であった）このコインを一回投げるという実験の期待値は、n回投げたときの結果の平均値からどのように得られるであろうか？そのとき、平均値はどのような分布になるであろうか？

ーー単に上の二項分布をnで割ればよい。

☆コイン投げの平均値の分布

deleted based on copyright concern.
unspecified quotation.

二項分布の標準偏差は $\sqrt{n}$ に比例なので、それを n で割った平均値の分布の標準偏差は $1/\sqrt{n}$ に比例となる。

例えば、nを10倍にすれば、分布の幅（平均値のバラつき）は $/\sqrt{}$ になる。

また、その平均値の分布はnが大きい場合には正規分布となる。

☆ 6面体サイコロの平均値の分布①

次に、6面体サイコロ、つまり立方体の各面に1から6までの数字が書いてあるサイコロをn回振って、サイコロを一回振ったときの出た目の平均値を求める。

正解としては、サイコロを1回振って出る目の期待値は

$$\mu = 1\cdot\frac{1}{6}+2\cdot\frac{1}{6}+3\cdot\frac{1}{6}+4\cdot\frac{1}{6}+5\cdot\frac{1}{6}+6\cdot\frac{1}{6}=3.5$$

標準偏差は、

$$\sigma = 1.71$$

となる。

では、n回サイコロを振り、その1回当たりの平均値を計算した場合、それは正しい期待値に対してどの程度ばらついているであろうか？

☆ 6面体サイコロの平均値の分布②

deleted based on copyright concern.
unspecified quotation.

図に見られるように、nを増やすことにより分布は正規分布に近づき、また分布の幅は計算するとやはり $1/\sqrt{n}$ に比例する。

☆ 改造した6面体サイコロの平均値の分布①

さらに特殊な例として、6面体サイコロを改造し、立方体の各面に1－6の代わりに1，1，1，2，3，3と書かれた改造サイコロを作成し、それをn回振って、サイコロを一回振ったときの出た目の平均値を求める。

正解としては、この改造サイコロを1回振って出る目の期待値は

$$\mu = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{3} = 1.83$$

標準偏差は、

$$\sigma = 0.89$$

となる。

では、同様にn回サイコロを振り、その1回当たりの平均値を計算した場合、その分布はどうなるであろうか？

☆ 改造した6面体サイコロの平均値の分布②

deleted based on copyright concern.
unspecified quotation.

その平均値が出た頻度

平均値

やはり図に見られるように、nを増やすことにより分布は正規分布に近づき、また分布の幅は計算すると $1/\sqrt{n}$ に比例する。

どんなコインでもサイコロでも、この性質は成り立つ？？

☆確率変数と確率分布①

以降の議論を簡単にするために、**確率変数**と**確率分布**という用語を導入する。

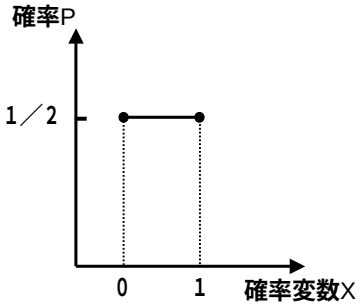
変数 X がm個の値 $X=x_1,x_2,\cdots,x_m$ を確率的にとり、そのそれぞれについて確率 $P(x_i):i=1,2\cdots m$ が与えられているとき、それを（離散）**確率変数**と呼ぶ。

例えば、最初のコイン投げの例では、確率変数 $X=x_1,x_2=0,1$ となり、そのときの対応する確率は

$$P(0)=\frac{1}{2}\ ,\ P(1)=\frac{1}{2}$$

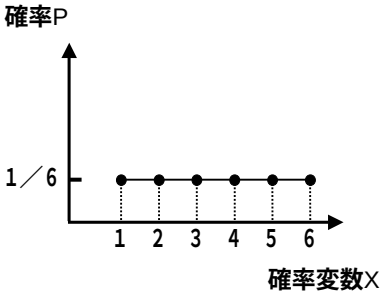
となる。

また、確率変数Xの取りうる全ての値と対応する確率がわかっているとき、「**Xの確率分布が与えられている**」といい、コイン投げの例では確率分布は右の図のようにまとめられる。

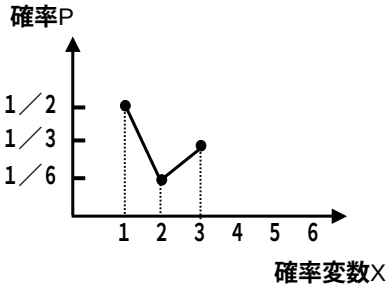


☆確率変数と確率分布②

同様に、6面体サイコロの確率分布は右のようになる。



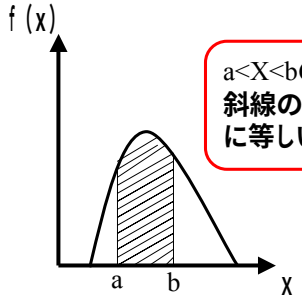
6面体サイコロを改造し、各面に1，1，1，2，3，3と書かれた改造サイコロの確率分布は右のようになる。



☆確率変数が連続的に変化する場合

確率変数Xが連続的に変化する場合は、変数Xがある値となる確率を定義することが出来ない。その場合には、以下のように**確率密度関数** $f(x)$を使い、変数Xが値aから値bの間にある確率を以下のように定義する。

$$P(a \le X \le b) = \int_a^b f(x)dx \ , \ \int_{-\infty}^{\infty} f(x)dx = 1$$



a<X<bの確率は、斜線の部分の面積に等しい

☆確率分布における期待値と標準偏差

確率変数 X がm個の値 $X = x_1, x_2, \cdots, x_m$ をとり、そのそれぞれについて確率 $P(x_i) : i = 1,2 \cdots m$ が与えられているとき、その確率変数の期待値 μ と標準偏差 σ は以下のように与えられる。

$$\mu = \sum_i x_i P_i = x_1 P_1 + x_2 P_2 + x_3 P_3 + \cdots$$

$$V[X] = \sum_i (x_i - \mu)^2 P_i = (x_1 - \mu)^2 P_1 + (x_2 - \mu)^2 P_2 + \cdots$$

$$\sigma = \sqrt{V[X]}$$

確率変数 Xが連続的に変化する場合は、確率密度関数を用いて、期待値と標準偏差は以下のように与えられる。

$$\mu = \int_{-\infty}^{\infty} xf(x)dx \ , \ \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$$

☆中心極限定理

確率変数Xが何らかの確率分布に従い、その確率分布が期待値 μ
標準偏差 σ を持つとする。

このとき、この確率分布から独立に得られたn個の確率変数 $X_1, X_2, \cdots, X_n$
の平均値

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}\sum_i X_i$$

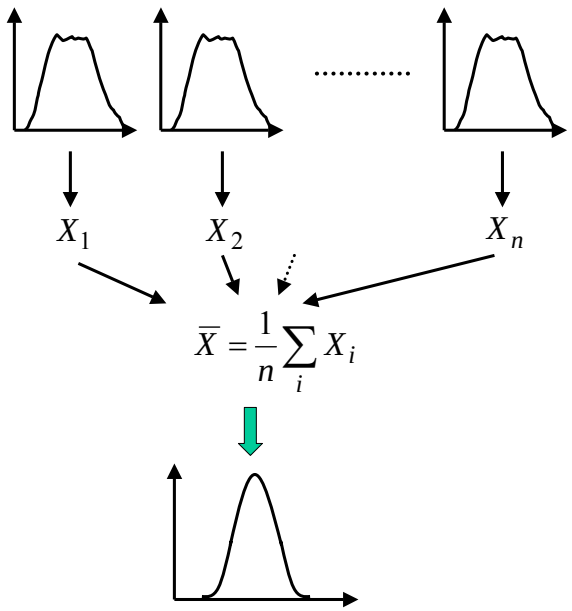
は、nが十分大きいときには平均 μ 、標準偏差 $\sigma' = \frac{\sigma}{\sqrt{n}}$ の正規分布

$$P(\overline{X}) = \frac{1}{\sqrt{2\pi}\sigma'} e^{-\frac{(\overline{X}-\mu)^2}{2(\sigma')^2}}$$

に従う。

（ただし、元の確率分布の期待値と標準偏差が発散しない場合）

☆中心極限定理のイメージ



期待値 μ 標準偏差 σ を
持つ同一の確率分布
（どんな分布でも良い！）

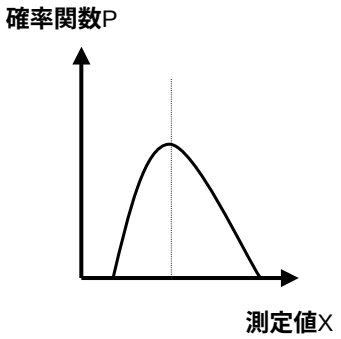
nが大きい場合には、
平均値は期待値 μ
標準偏差 $\sigma/\sqrt{n}$ の
正規分布に従う

☆母集団と標本集団①

次に、中心極限定理の考え方を基にして、誤差を含んだ実験結果を平均するということはどういうことかを考えてみる。

実験の結果はさまざまな理由により誤差を含み、その測定値は一定ではなく、ある幅をもったものとなる。

そのため、**測定値は何らかの確率変数にならざるを得ない**。

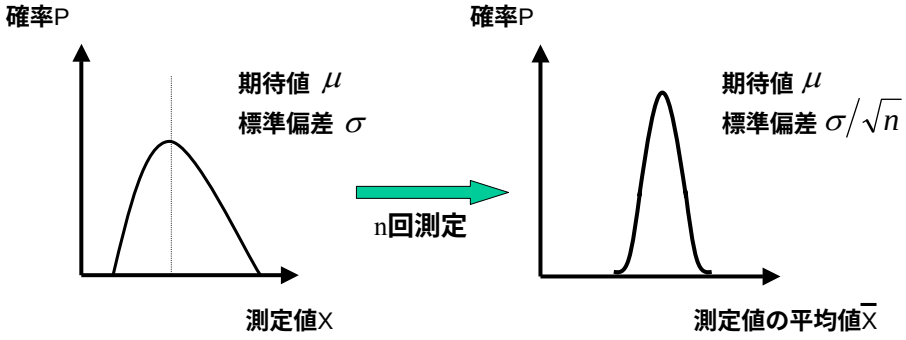


この測定値の確率分布を正確に決定するためには、無限個のデータが必要となる（そのような無限個のデータの集団を**母集団**と呼ぶ）。

しかしながら、我々は無限個のデータを得ることが出来ないので、母集団から有限個のデータをサンプルし、それを基に母集団の性質（つまりは確率分布の性質）を推定する必要がある（このようなサンプルされた有限個のデータを**標本集団**と呼ぶ）。

☆母集団と標本集団②

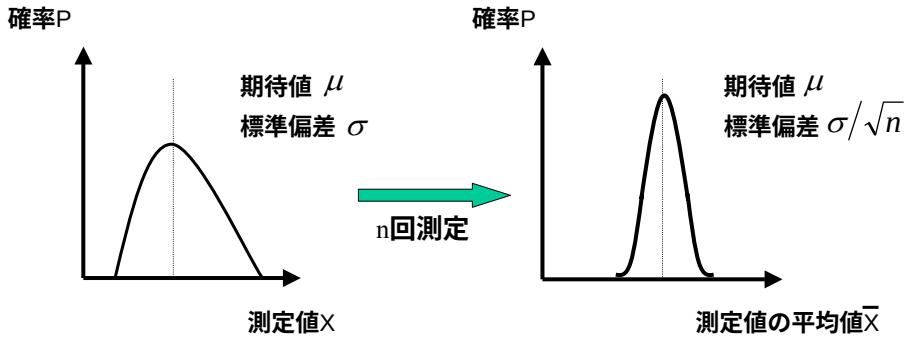
母集団の確率分布が期待値 μ 標準偏差 σ を持つとする。
この母集団からn個の測定値 $X_1, X_2, \cdots, X_n$ を得て、それの平均値 $\overline{X}$ を求めると、平均値 $\overline{X}$ は期待値 μ、 標準偏差 $\sigma/\sqrt{n}$ の正規分布に従う。



☆母集団の期待値の推定（の初歩）

では、実際にn回の測定を行って、平均値 $\overline{X}$ が得られたとする。

そのとき、母集団の標準偏差 σ が既知であるとすると、
平均値 $\overline{X}$ が標準偏差 $\sigma/\sqrt{n}$ の正規分布に従うことを利用して、
母集団の期待値がどの範囲にあるか推測することができる。



☆本日のまとめ

- 実験において、同じ条件で複数回測定するということは、その有限個からなる標本集団から、背後にある母集団の性質（例えば、期待値）を推測するためである。
- n回の測定の平均値は、母集団の期待値を中心とし、分布の幅が $1/\sqrt{n}$ に比例する正規分布に従う（中心極限定理）。つまり、平均値のバラつきの幅は測定回数を10倍にすると3分の1程度になる。
- 一言で言えば、中心極限定理とは、ランダムなものを足し合わせると何でも正規分布になってしまうことを意味している。