■モザイク図

モザイク図とは、クロス集計表を面積により視覚的に表現した図。[1]
あまり、科学技術の世界ではあまり使われない。（統計の試験ではよく出る）

前回も使った血液型の集計表

	1：A 型	2：B 型	3：O 型	4：AB 型	小計
1：男性	5	2	2	1	10
2：女性	3	2	4	1	10
小計	8	4	6	2	合計：20

◆行幅を構成比に合わせたクロス集計表

	1：A 型	2：B 型	3：O 型	4：AB 型
1：男性	5	2	2	1
2：女性	3	2	4	1

◆列の高さを構成比に合わせたクロス集計表

	1：A 型	2：B 型	3：O 型	4：AB 型
1：男性	5	2	2	1
2：女性	3	2	4	1

[1] モザイク図を Office ツールで作るのは少し面倒

■リスク比とオッズ比

どちらも、2×2 のクラス集計表で、2 つの要因の関連性を図るための指標。

	変数 Y-1	変数 Y-2
変数 X-1	A	B
変数 X-2	C	D

リスク比と**オッズ比**を求める式

リスク比（相対危険度）：

$$\frac{A}{A+B} \div \frac{C}{C+D}$$

オッズ比：

$$\frac{A}{B} \div \frac{C}{D}$$

◆リスク比・オッズ比の実習

喫煙と肺がんの関係のクロス集計表からリスク比とオッズ比を求めてみる

	肺がん Yes	肺がん No
喫煙あり	20	80
喫煙なし	10	90

この結果、リスク比は「2」になる。
つまり、喫煙者は非喫煙者の 2 倍のリスクで肺がんになる。こうした結果を導けるのが「リスク比」という名前の由縁。

オッズ比を求めると、「2．25」になる。
この意味は、喫煙者と非喫煙者との間で、肺がんを患っていない人に対する肺がん患者の割合を比べている。喫煙者の方が 2.25 倍高いと言える。

◆リスク比もオッズ比も 1 に近づく意味

	肺がん Yes	肺がん No
喫煙あり	20	80
喫煙なし	20	80

クラス集計表から喫煙者も非喫煙者も同じ割合で肺がんに罹患する。
こうしたデータでは、リスク比もオッズ比も「1」となる。2 つの変数の値が近いほど 1 に近づく。

◆層別分析

1：好む	2：好まない	小計
600	400	1,000

このように 1 変数だけの表を**単純集計表**といいます。

このデータと同時に他のカテゴリをつける、たとえば「性別」、「地域」などの要素を加える分析を「**層別分析**」と呼びます。

	1：好む	2：好まない	小計
1：関東	250	250	500
2：関西	350	150	500
	600	400	1,000

■確立

　確立とは、不確実な事象の起こりやすさを数値で表したもの。
確立を「P」で表し、事象 A の確立を $P(A)$ と表す。
　確立 P は 0 から 1 までの数で表される。数が大きいほど、起こりやすいことを示す。
さらに、全ての事象の確率を足し合わせると「1」になる。

◆確立の実例

　袋の中に「赤玉 4 つ」と「白玉 1 つ」が入っている。中を見ずランダムに 1 つ取り出すことを考える。
このとき、取り出した玉の色は「赤」であるか「白」であるかの 2 つ。
　高校数学までの確率の範囲では、赤玉が出る場合の数は 4，白玉が出る場合の数は 1。全事象の起こる場合の数は 5。したがって、赤玉、白玉のでるそれぞれの確立は、

$$P(\text{赤玉})=4/5、P(\text{白玉})=1/5$$

となる。
　この場合、取り出す玉は「等しい確立で取り出す」という仮定が含まれている。

◆確率変数

　変数 X＝［赤玉、白玉］と置くと、

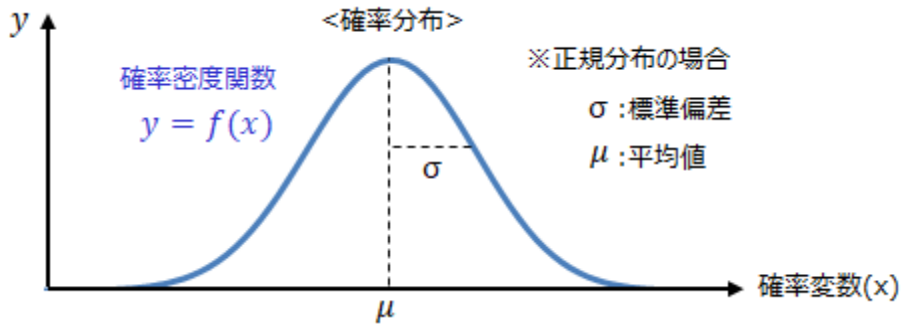
$$P(X=\text{赤玉}=4/5、\ P(X=\text{白玉})=1/5$$

と表すことができる。
　この X のように確率的に変動する変数を**確率変動変数**とよぶ。
それに対して、確率変数が実際に取る値（赤玉か白玉か）を実現値とよぶ。
　この場合、**離散型確率変数**といえる。

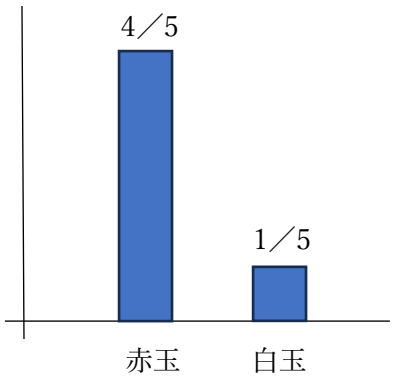
◆確率分布

確率分布とは、横軸に「確率変数」、縦軸に「確率変数の起こりやすさ」を表した分布。

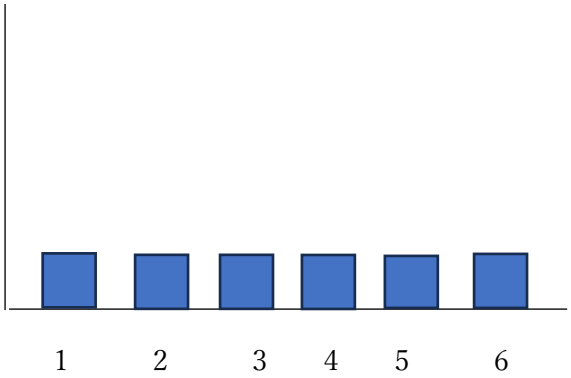
サイコロのような「離散型の確率分布」の場合、縦軸は「確率」そのものを表し、横軸は「確率変数」を表している。



カテゴリ変数の場合



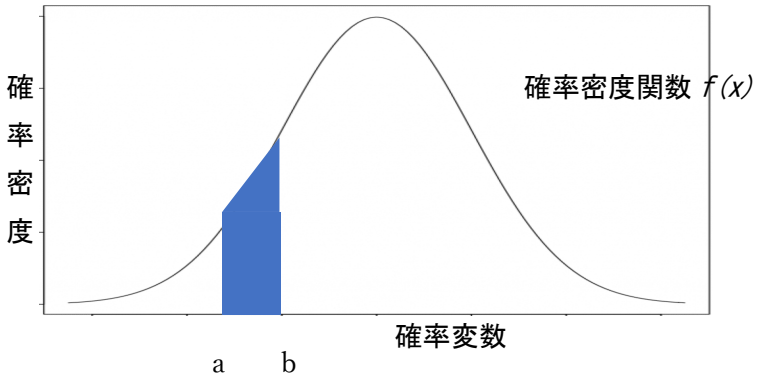
サイコロなど離散型の量的変数の場合



確率変数が連続的であると確率変数が実数値となり、小数点以下が続く値になる。

そこで、連続型確率変数の場合には、値に幅を持たせて確率を求める。このための関数を「**確率密度関数**」と呼びます。この関数は、確率そのものではなく、確率の起こりやすさを表わす数値。

確率密度の関数の積分を計算して、X 軸と確率密度関数で囲まれた面積を、つまり、この面積が確率になる。



a から b の範囲の値が現われる確率は、その区間の面積に対応する

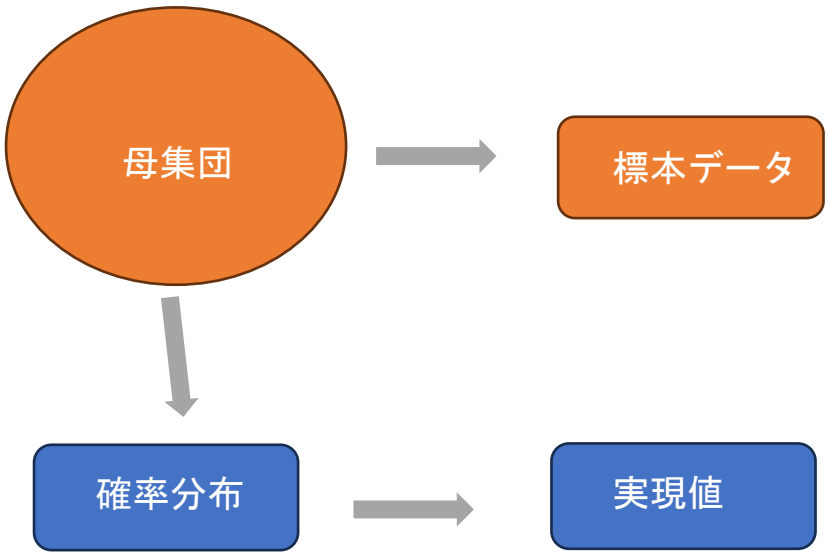
$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

◆推測統計と確率分布

推測統計では、母集団の一部である標本から母集団の性質を推測する。しかし、母集団は直接観測できないし、標本から推測することは困難に思える。

そこで、現実世界の母集団を数学の世界の確率分布と仮定して、標本のデータはその確率分布の中か空生成された「実現値」であるという仮定のもとに分析をする。

したがって、「母集団と標本データ」という扱いづらい対象が、「確率分布とその実現値」という数学的にとらえる対象に置き換えられた。



データは、ある確率分布から得られた実現値であると考える

母集団から抽出された標本を、ある数学的な確率分布から生成された値だと仮定することによって分析を進めていくことが可能になる。また、統計的推測とは、データから生成元がどのような確率分布であるかを推測することになる。

■期待値

期待値とは、1 回の試行で得られる値の平均値のことをいう。
つまり、得られる全ての値とそれが起こる確率の積を足したもの。

◆サイコロを 1 回投げたときの期待値

サイコロを 1 回投げたときに出る目と確率は、

出る目	1	2	3	4	5	6
確率	1/6	1/6	1/6	1/6	1/6	1/6

この期待値の計算は、

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

つまり、サイコロを 1 回投げてでる目の平均は 3.5 になる。

◆コイントスで賭けると

コインの表がでたら 1000 円、裏がでたら-1000 円（千円払う）とする。
この場合の確率は、

コイン	表	裏
もらえるお金	1000 円	-1000 円
確率（p）	1/2	1/2

これを計算すると

$$1000\times\frac{1}{2}+(-1000)\times\frac{1}{2}=0$$

したがって、期待値は 0 円になるので、この加計には参加しない方が良い。