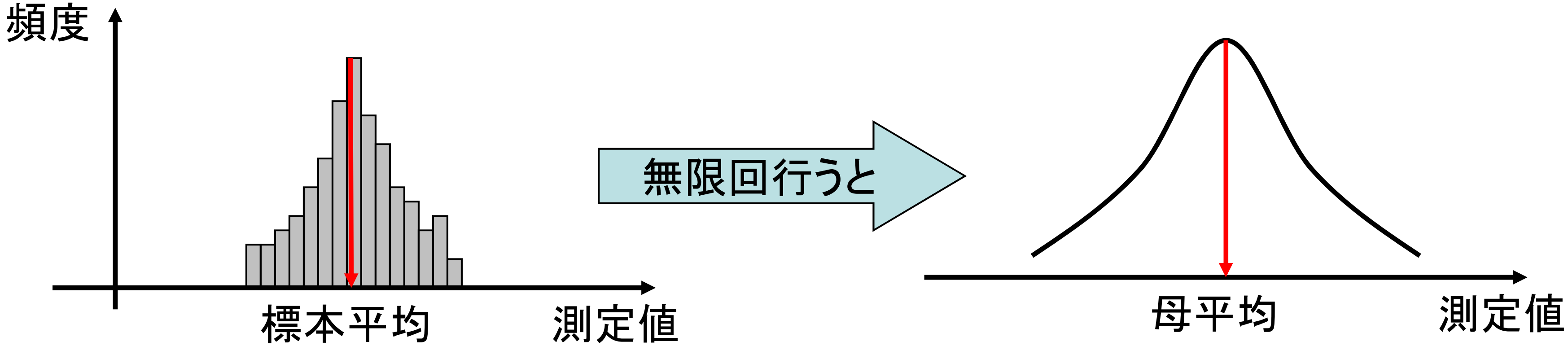
不確かさ評価に必要な統計的基礎

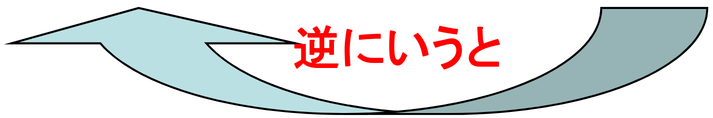
産業技術総合研究所　計量標準総合センター
田中秀幸

測定と統計について



測定を行えばこのような
ヒストグラムが得られる。

無限回測定するとこのような測定の母
集団が得られる。

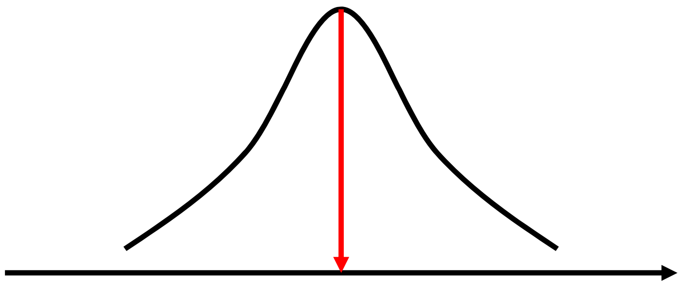


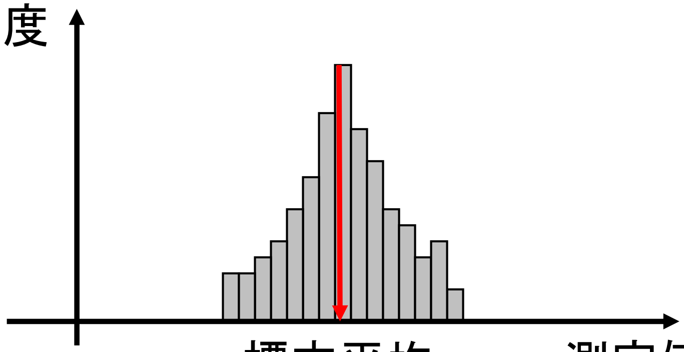
<u>**測定をしようと決めたときに，すでに母集団の形は決まっている。**</u>
<u>**測定を行えばその母集団から値をひとつ取り出すということになる。**</u>

測定とは，その測定を無限回繰り返したとき得られる
母集団からのサンプリングである。

母集団と標本について(1)

私たちは測定することによってなにを知りたいのだろう？





本当ならこの母集団の性質を知りたい。しかし，母集団の性質も理想的な量であるので知ることはできない。

そこで，無限回測定する事は無理なので有限回の測定によって母集団の性質を推定する。

サンプルの性質を知りたいわけではない

一番知りたいもの	***母集団の性質***

統計の重要な目的は，この母平均の推定値とその推定された値がどの程度正しいのかということ知ることである。

用いる文字について

標本を表すもの：アルファベット
母集団の性質を表すもの（母数）：ギリシャ文字
ギリシャ文字の上に「^」（ハット）が付くもの：推定値

i番目の測定結果：　x_i
標本平均：　$\overline{x}$
母平均：　μ

標本平均を母平均の推定値とする：

$$\hat{\mu} = \overline{x}$$

連続分布と離散分布

離散分布･･･飛び飛びの値しか取らない要素からなる分布
例：サイコロの目の値，個数，人数････

連続分布･･･離散分布とは違い，ある区間内に値が存在しているとしても，その区間内で無限個のデータが含まれている要素からなる分布
例：測定結果

離散分布は概念や統計処理において理解しやすいが，連続分布は無限の概念を考える必要があり，理解しにくい。
ただし，測定結果は通常連続分布であるので,連続分布における統計処理について知る必要がある。

期待値とは

- 期待値･･･理想的にはこの値になるという値

サイコロは1－6の面を一つずつ持っている.

また，その6つの面がでる確率はそれぞれ1/6.

よって期待値は,

$$1\cdot\frac{1}{6}+2\cdot\frac{1}{6}+3\cdot\frac{1}{6}+4\cdot\frac{1}{6}+5\cdot\frac{1}{6}+6\cdot\frac{1}{6}$$
$$=\frac{1}{6}+\frac{2}{6}+\frac{3}{6}+\frac{4}{6}+\frac{5}{6}+\frac{6}{6}=\frac{21}{6}=3.5$$

期待値とは

- 一般的に書くと

$$E(x) = x_1 P(x_1) + x_2 P(x_2) + \cdots + x_n P(x_n)$$

$$E(x) = \sum_{i=1}^{n} x_i P(x_i)$$

期待値は，$E(x)$で表す

$P(x_i)$のことを確率関数という

期待値の性質

測定値xの母平均をμとすると，	$\mu = E(x)$
cが定数のとき，	$E(c) = c \quad E(cx) = cE(x)$
$x,\ y$が確率変数のとき，	$E(x \pm y) = E(x) \pm E(y)$
$x,\ y$が確率変数で互いに独立のとき，	$E(xy) = E(x)E(y)$

期待値の性質

cが定数のとき，　$E(c) = c$　　$E(cx) = cE(x)$

$$E(c) = c$$

cは確率的に値が変わらないので，
いつでもcとなる。

$$E(cx) = cE(x)$$

2,4,6,8,10,12のサイコロを考えると，期待値は(2+4+6+8+10+12)/6=7となる。
これは，普通のサイコロの期待値3.5の2倍である。

期待値の性質

$x,\ y$が確率変数のとき，$E(x \pm y) = E(x) \pm E(y)$

	1	2	3	4	5	6	和
1	2	3	4	5	6	7	27
2	3	4	5	6	7	8	33
3	4	5	6	7	8	9	39
4	5	6	7	8	9	10	45
5	6	7	8	9	10	11	51
6	7	8	9	10	11	12	57

全和:252
$E(x+y)=$
$252 \div 36=7$

$E(x)+E(y)=$
$3.5+3.5=7$

二つのサイコロを同時に振って出た目を足した値について考える。二つのサイコロの出目の和は2と12の間になり，その中心の値（期待値）は7である。この値は，一つだけサイコロを振ったときの期待値3.5を2つ足し算したときと同じである。

期待値の性質

x, *y*が確率変数で互いに独立のとき，

$$E(xy) = E(x)E(y)$$

	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	4	6	8	10	12
3	3	6	9	12	15	18
4	4	8	12	16	20	24
5	5	10	15	20	25	30
6	6	12	18	24	30	36
和	21	42	63	84	105	126

全和：441
$E(xy)=$
$441 \div 36 = 12.25$

$E(x)E(y)=$
$3.5 \times 3.5 = 12.25$

互いに独立：確率変数*x*と*y*が互いに影響し合うことがなく，全く別個に値が決定する，ということを意味している。

標本平均の期待値

標本平均: $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$ の期待値を考える。

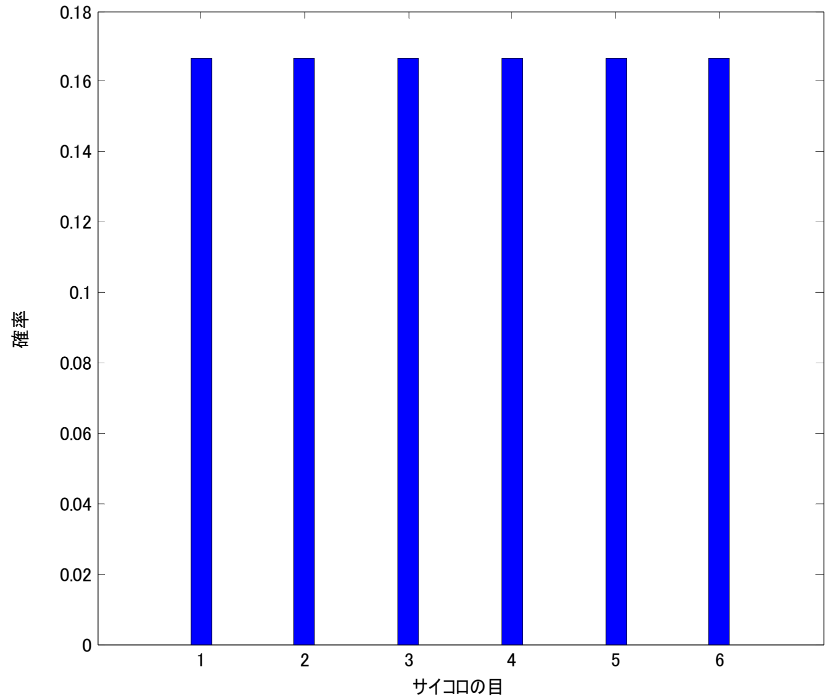
$$E(\bar{x}) = E\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{1}{n}E(x_1 + \cdots + x_n)$$

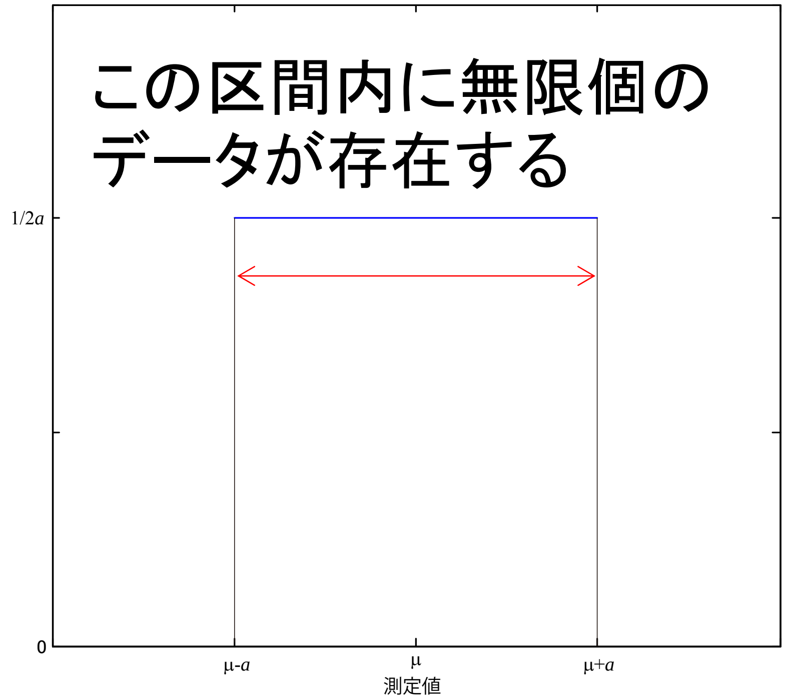
$$= \frac{1}{n}\left[E(x_1) + \cdots + E(x_n)\right] = \frac{1}{n}\sum_{i=1}^{n} E(x) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}\cdot n\mu$$

$$E(\bar{x}) = \mu$$

標本平均は母平均の推定値として何らかのかたよりを持たない。このようにかたより無く母数の推定が行うことができる推定量のことを**不偏推定量**という。

連続分布への拡張





離散分布

ある値を取るときの確率を求めることができる。

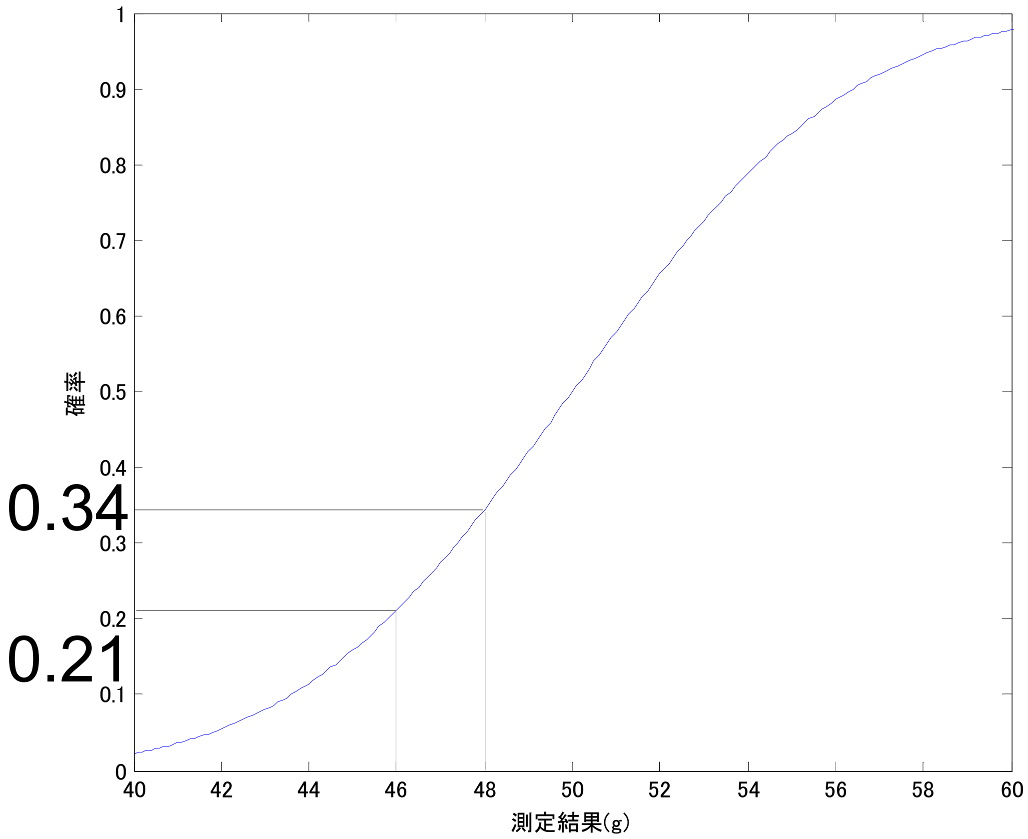
連続分布

ある値を取るときの確率を求めることができない。

無限個からある要素一つがサンプリングされる確率は, $\frac{1}{\infty} = 0$

連続分布への拡張

では，連続分布のときは確率が計算できないのか？



製品の質量の分布（累積分布関数）

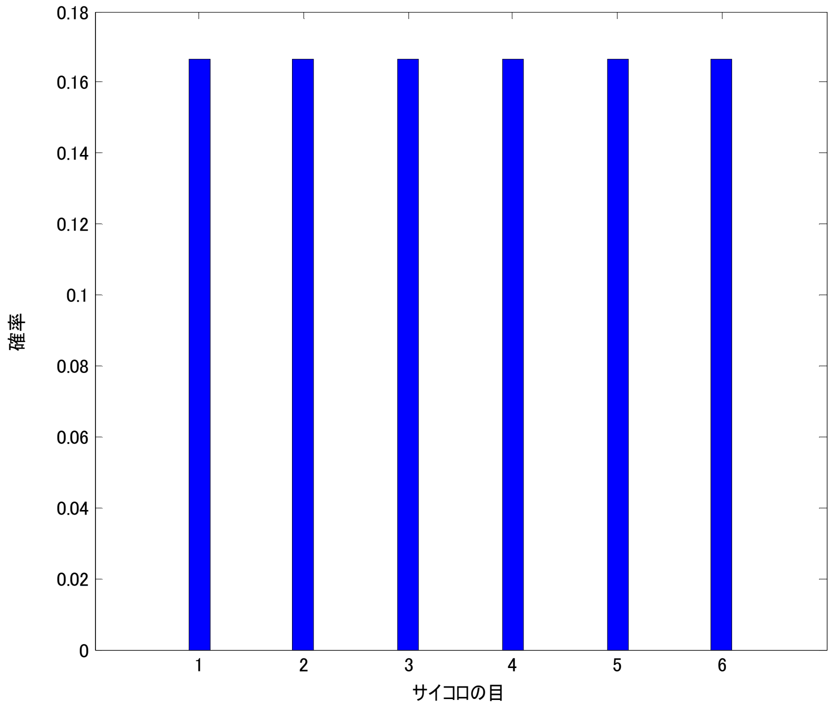
→ある値になるときの確率が計算できないだけで，サンプリングされた値が，ある範囲内に含まれる確率であれば計算できる。

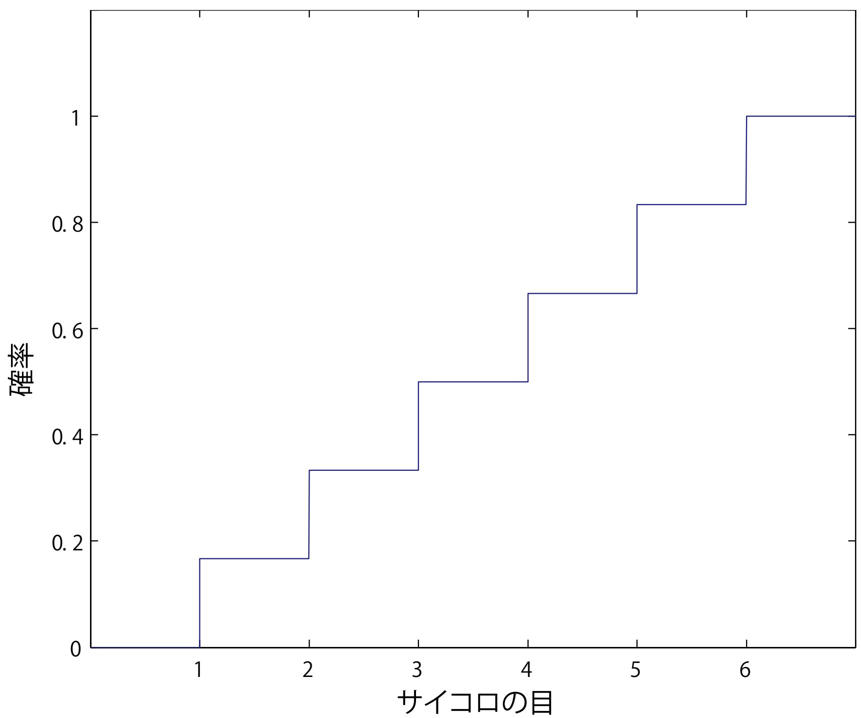
測定結果が46gから48gの間に入る確率は，

$$0.34 - 0.21 = 0.13$$

左図で示したグラフを表す関数を**累積分布関数**または**分布関数**（Cumulative Distribution Function: **CDF**）といい，この分布関数は$F(x)$で表す。

離散分布の累積分布関数





確率関数と累積分布関数の関係を考えると，確率関数では，1-6の目がすべて1/6の確率で出ており，累積分布関数では，目が1増える毎に1/6ずつ確率が増えることが分かる。つまり，確率関数は累積分布関数の傾きを表していることになる。

確率密度関数

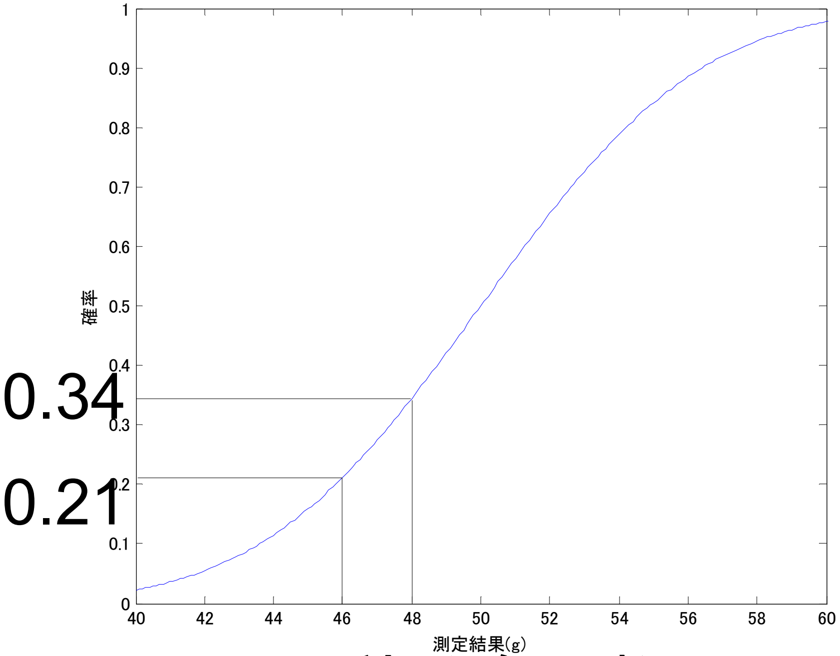
離散分布
確率関数は累積分布関数の傾きを表している

↓

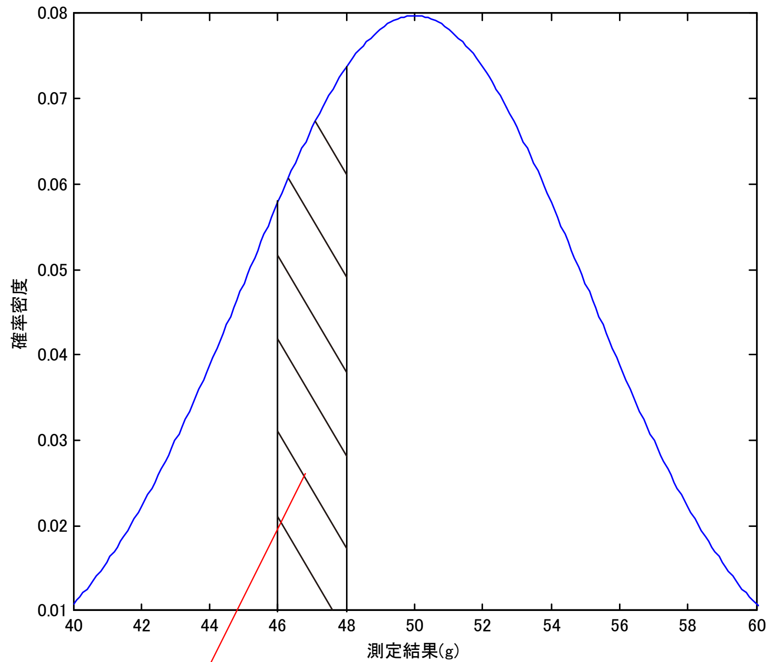
連続分布
累積分布関数の傾きを表した関数（つまり累積分布関数を微分したもの）が，連続分布における確率関数のようなものになる

この連続分布における確率関数に相当するものを
確率密度関数（Probability Density Function: **PDF**）という

確率密度関数



累積分布関数



確率密度関数

測定結果が46gから48gの間に入る確率は, $0.34-0.21=0.13$

ここの面積を求めると0.13

確率密度関数のグラフは直感的に母集団の分布を示しているので, 一般的に確率分布を示すときには確率密度関数が使われる。

連続分布の期待値

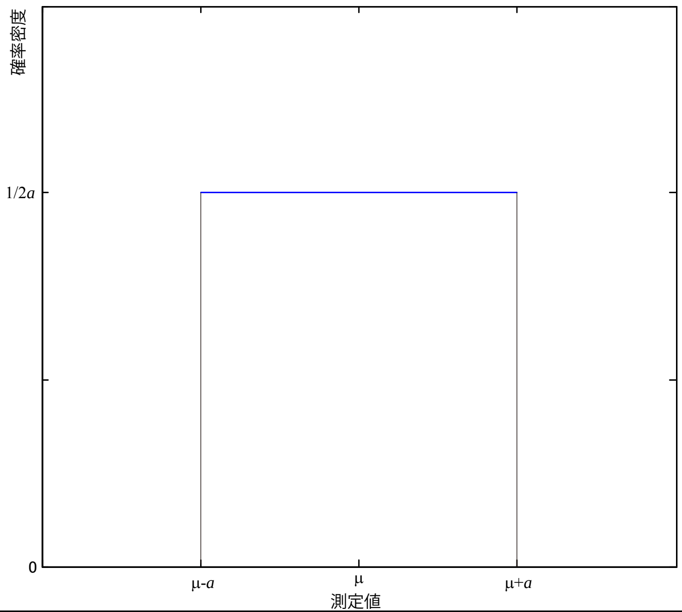
離散分布

$$E(x)=\sum_{i=1}^{\infty}x_iP(x_i)$$

連続分布

$$E(x)=\int_{-\infty}^{\infty}xf(x)dx$$

測定結果に(確率関数)(確率密度関数)を掛けて和を取る。



矩形分布

$$f(x)=\begin{cases}\dfrac{1}{2a} & (\mu-a\le x\le \mu+a)\\ 0 & (\text{それ以外})\end{cases}$$

$$E(x)=\int_{-\infty}^{\infty}xf(x)dx=\int_{\mu-a}^{\mu+a}x\frac{1}{2a}dx=\frac{1}{2a}\int_{\mu-a}^{\mu+a}xdx$$

$$=\frac{1}{2a}\left[\frac{1}{2}x^2\right]_{\mu-a}^{\mu+a}=\frac{1}{2a}\left\{\frac{1}{2}(\mu+a)^2-\frac{1}{2}(\mu-a)^2\right\}$$

$$=\mu$$

母集団のばらつき

個々のデータが持つばらつきの大きさ: $\delta = x - \mu$　　　δ: 偏差

平均的なばらつきを知るために偏差の期待値を求める

$$E(\delta) = E(x) - \mu = \mu - \mu = 0$$

xは母平均より大きな値，小さな値の両方取り得る可能性があるから

δの大きさに着目する必要がある ⇒ δを二乗する

$$E\left(\delta^2\right) = E\left[(x-\mu)^2\right]$$

この$E(\delta^2)$のことを分散といい$V(x)$で表す。 $V(x) = E\left[(x-\mu)^2\right]$

母分散について

$$V(x)=E\left[(x-\mu)^2\right]$$

$V(x)$はxのばらつきを表すパラメータであるので，
これを母分散といいσ^2で表される。

なぜσに二乗が付くのかというと，例えば，測定結果が質量の[g]で表されていたとすると，この母分散の単位は偏差を二乗しているので，[g^2]となるからである。また，二乗が付かないσ単体で表される母数のことを母標準偏差という。

分散の別の表し方

$$V(x)=E\left[(x-\mu)^2\right]=E\left(x^2+\mu^2-2\mu x\right)$$

$$=E\left(x^2\right)+\mu^2-2\mu E(x)=E\left(x^2\right)-\mu^2$$

$$=E\left(x^2\right)-\left[E(x)\right]^2$$

母分散の計算法

離散分布

$$V(x)=E\left[(x-\mu)^2\right]=\sum_{i=1}^{\infty}(x_i-\mu)^2 P(x_i)$$

連続分布

$$V(x)=E\left[(x-\mu)^2\right]=\int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx$$

(測定結果-母平均)^{2}に（確率関数）（確率密度関数）を掛けて和を取る。

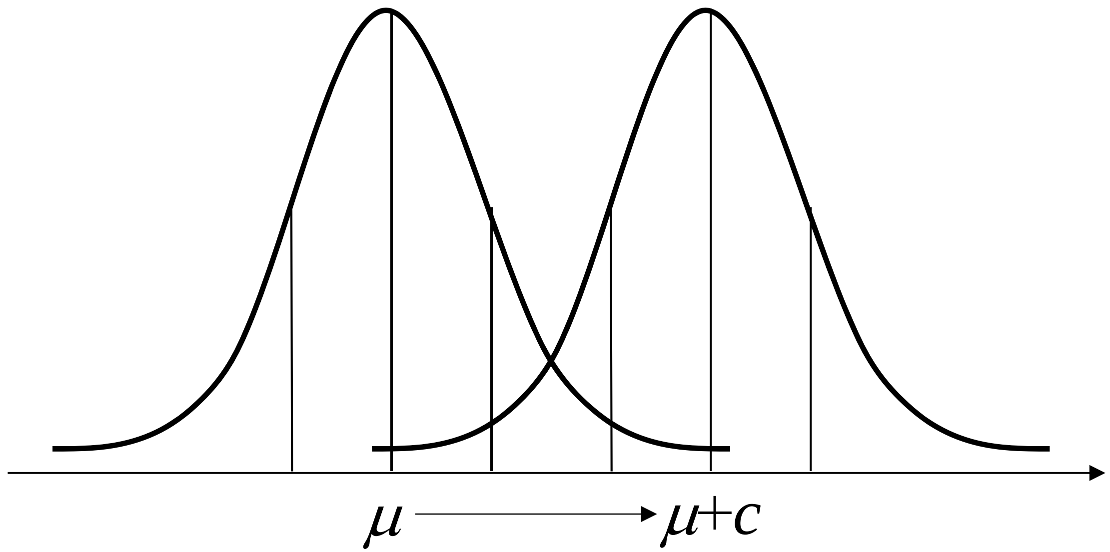
cが定数のとき, $V(x+c)=V(x)$ $V(cx)=c^2V(x)$

x, yが確率変数で互いに独立のとき, $V(x\pm y)=V(x)+V(y)$

分散の性質

cが定数のとき， $V(x+c)=V(x)$

xがcだけ平行移動したときのばらつきを考える。確率分布が平行移動しただけなら，分布の形自体は変わらないことからばらつきの大きさも変わらない。

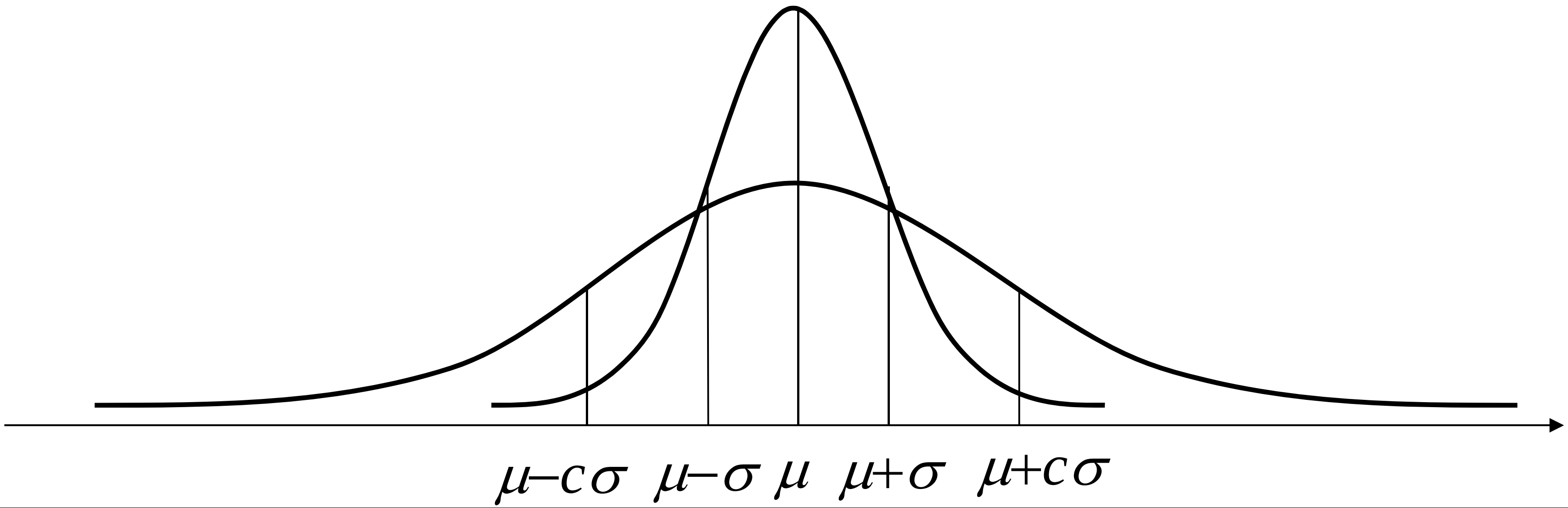


分散の性質

cが定数のとき， $V(cx)=c^2V(x)$

確率変数がc倍されるのであるから，確率分布のばらつきはc倍される。ただし，母分散の単位は測定対象量の単位の二乗となっているので，母集団のばらつきがc倍されるのであれば，分散はc^2倍される。

$$
\begin{aligned}
V(cx) &= E\left[\left\{c\left(x-\mu\right)\right\}^2\right] \\
&= E\left[c^2\left(x-\mu\right)^2\right] \\
&= c^2E\left[\left(x-\mu\right)^2\right] = c^2V(x)
\end{aligned}
$$



分散の性質

x, yが確率変数で互いに独立のとき，　$V(x \pm y) = V(x) + V(y)$

$$V(x \pm y) = E\left[\left\{(x-\mu_x) \pm (y-\mu_y)\right\}^2\right]$$
$$= E\left[(x-\mu_x)^2 + (y-\mu_y)^2 \pm 2(x-\mu_x)(y-\mu_y)\right]$$
$$= E\left[(x-\mu_x)^2\right] + E\left[(y-\mu_y)^2\right] \pm 2E\left[(x-\mu_x)(y-\mu_y)\right]$$

$E\left[(x-\mu_x)(y-\mu_y)\right]$は確率変数$x, y$が独立であるとき，

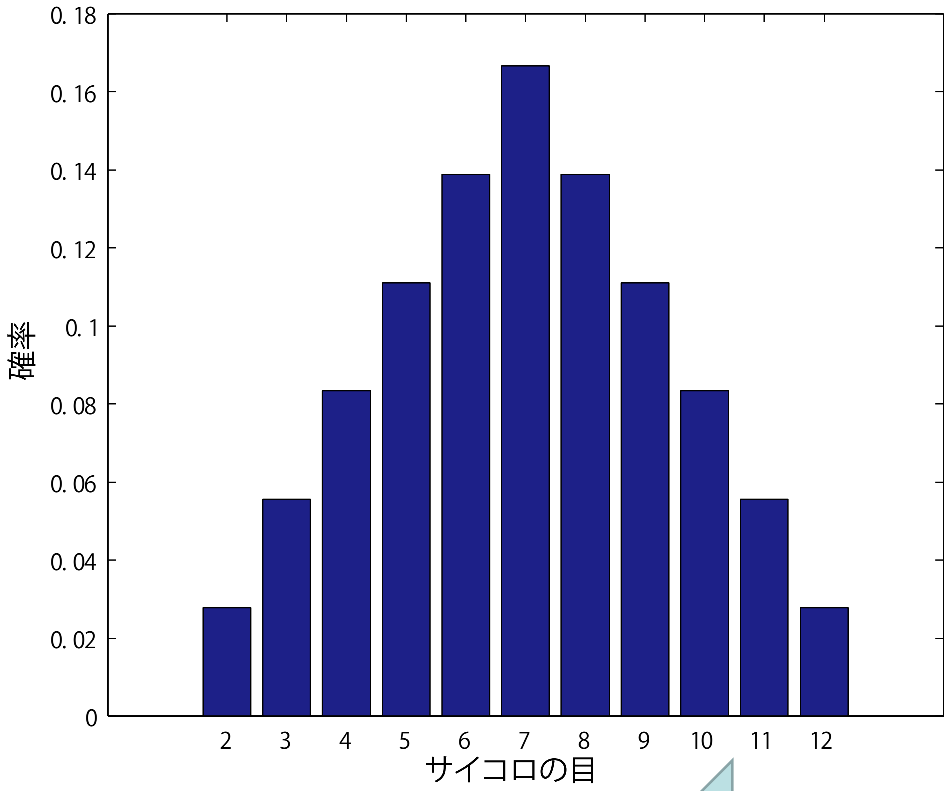
$$E\left[(x-\mu_x)(y-\mu_y)\right] = 0$$

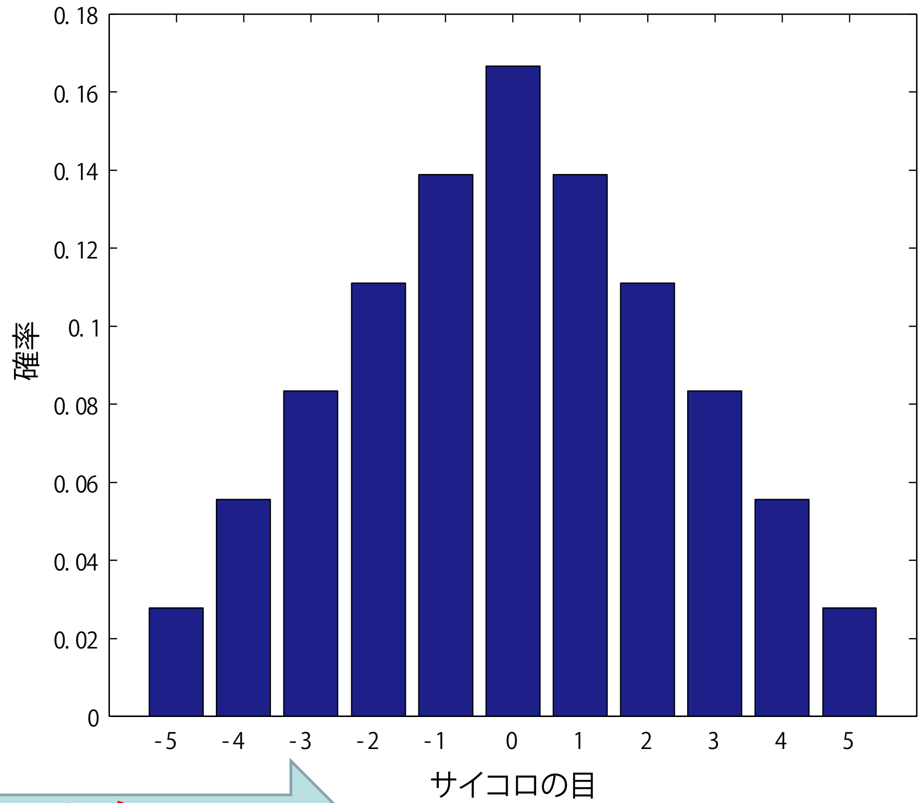
$$V(x \pm y) = E\left[(x-\mu_x)^2\right] + E\left[(y-\mu_y)^2\right]$$
$$= V(x) + V(y)$$

分散の性質

x, yが確率変数で
互いに独立のとき,

$$V(x \pm y) = V(x) + V(y)$$





和 　同じ分布　 差

母分散と標本分散

母分散は母数なので実際には知ることができない。よって，母分散を標本分散によって推定することを考える。

残差：　$d_i = x_i - \overline{x}$　　（母平均を標本平均で代用）

期待値を求める代わりに標本平均を求めて代用する。

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

s^2が標本分散である。この標本分散によって，母分散を推定する。

標本分散と不偏推定量

標本分散は母分散の不偏推定量であるかを考える。

不偏推定量であればもちろん，$E\left(s^2\right)=V\left(x\right)=\sigma^2$ が成立する。

$$E\left(s^2\right)=E\left[\frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}{n}\right]=\frac{1}{n}E\left[\sum_{i=1}^{n}\left(x_i^2-2x_i\overline{x}+\overline{x}^2\right)\right]$$

$$=\frac{1}{n}E\left[\sum_{i=1}^{n}x_i^2-2\overline{x}\sum_{i=1}^{n}x_i+\overline{x}^2\sum_{i=1}^{n}1\right]=\frac{1}{n}E\left[\sum_{i=1}^{n}\left(x_i^2\right)-2\overline{x}\cdot n\overline{x}+n\overline{x}^2\right]$$

$$=\frac{1}{n}E\left[\sum_{i=1}^{n}\left(x_i^2\right)-n\overline{x}^2\right]=\frac{1}{n}\sum_{i=1}^{n}E\left(x_i^2\right)-E\left(\overline{x}^2\right)$$

$E\left(x_i^2\right)=E\left(x^2\right)$ と考えてよいので，$E\left(s^2\right)=\frac{1}{n}\sum_{i=1}^{n}E\left(x^2\right)-E\left(\overline{x}^2\right)$

標本分散と不偏推定量

ここで, $E\left(x^2\right)$と$E\left(\overline{x}^2\right)$について考える。

$$V(x)=E\left[(x-\mu)^2\right]=E\left(x^2+\mu^2-2\mu x\right)$$
$$=E\left(x^2\right)+\mu^2-2\mu E(x)$$
$$=E\left(x^2\right)-\mu^2$$

$$E\left(x^2\right)=\sigma^2+\mu^2$$

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

$$V(\overline{x})=E\left(\overline{x}^2\right)-\mu^2$$
$$\frac{V(x)}{n}=E\left(\overline{x}^2\right)-\mu^2$$
$$E\left(\overline{x}^2\right)=\frac{\sigma^2}{n}+\mu$$

標本分散と不偏推定量

両式を代入

$$E\left(s^2\right)=\frac{1}{n}\sum_{i=1}^{n}E\left(x_i^2\right)-E\left(\overline{x}^2\right)$$
$$=\frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2+\mu^2\right)-\left(\frac{\sigma^2}{n}+\mu^2\right)$$
$$=\frac{1}{n}\cdot n\left(\sigma^2+\mu^2\right)-\left(\frac{\sigma^2}{n}+\mu^2\right)$$
$$=\sigma^2-\frac{\sigma^2}{n}$$
$$=\frac{n-1}{n}\sigma^2$$

つまり，$E\left(s^2\right)=V\left(x\right)$は成立しない。

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

上式は母分散を$(n\text{-}1)/n$倍された値，すなわち母分散より少し小さな値を推定していることとなる。よって不偏推定量であるとは言えない。

標本分散と不偏推定量

標本分散を不偏推定量にするためには，先程の式を$n/(n\text{-}1)$倍すればよい

$$s^2(x) = \frac{n}{n-1} \cdot \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1}$$

上式で表した標本分散のことを不偏分散という。通常測定データの処理には不偏分散を用いるので，これ以降の本解説内では，$s^2(x)$はすべて上式の不偏分散を表し，nで割った分散の意味では用いない。

直感的な解説

例：ある製品の質量測定(g)

x_1	x_2	x_3	x_4	x_5
87.5	86.2	90.1	88.4	87.0

標本平均：$\bar{x} = \dfrac{87.5+86.2+90.1+88.4+87.0}{5} = 87.84\ \ \mathrm{g}$

標本平均からの距離（残差）単位：g

87.5-87.84=-0.34
86.2-87.84=-1.64
90.1-87.84=2.26
88.4-87.84=0.56
87.0-87.84=-0.84

(標本平均からの距離)^2
単位：g^2

0.1156
2.6896
5.1076
0.3136
0.7056

残差の二乗和
単位：g^2

8.9320

データの個数-1
（自由度）で割る
単位：g^2

標本分散

2.233

平方根
単位：g

標本標準偏差

1.494

直感的な解説

つまり，それぞれ求められる残差の中で1つだけデータが分からなくなってしまっても，それは計算すれば求めることができる。つまり，残差はn個存在するが，本当に意味のある残差はn-1個となっているということである。

これは，残差を算出するために標本平均を使っていることが原因である。データはn個存在するが，残差を算出するために標本平均を用いるので，その標本平均にデータが持っている情報のうちデーター つ分が費やされているということである。

この本当に意味のあるデータの個数のことを**自由度**という。よって，標本分散を不偏推定量にするためにはデータの個数で残差の二乗和を割るのではなく，残差の二乗和を自由度で割らなくてはいけないのである。

標本平均のばらつき

母分散は測定データのばらつきを表す母数であるが，通常複数回の測定を行った場合，測定結果として報告する値は，その複数回測定したデータの平均値，つまり標本平均である。

↓

サイコロ5回振った平均値というものを考えてみると，あるときには平均値が3.2，あるときには2.4，もしかすると5回連続1がでて平均値が1となってしまう場合もあるだろう。このように，標本平均もばらつきを持つ値である。

標本平均のばらつきを考える必要がある

標本平均の母分散

標本平均の母分散は， $V\left(\overline{x}\right)=V\left[\dfrac{\sum_{i=1}^{n} x_i}{n}\right]$ と表すことができる。

$$V\left(\overline{x}\right)=\frac{1}{n^2}V\left(x_1+\cdots+x_n\right)$$

測定結果が独立であれば，

$$V\left(\overline{x}\right)=\frac{1}{n^2}V\left(x_1+\cdots+x_n\right)$$
$$=\frac{1}{n^2}\left[V\left(x_1\right)+V\left(x_2\right)+\cdots+V\left(x_n\right)\right]$$

標本平均の母分散

$$V\left(\overline{x}\right)=\frac{1}{n^2}\left[V\left(x_1\right)+V\left(x_2\right)+\cdots+V\left(x_n\right)\right]$$

$$=\frac{1}{n^2}\left[V\left(x\right)+V\left(x\right)+\cdots+V\left(x\right)\right]=\frac{1}{n^2}\cdot nV\left(x\right)$$

$$=\frac{V\left(x\right)}{n}=\frac{\sigma^2}{n}$$

標本平均の母分散は測定データの母分散の$1/n$となる。

サイコロを考えても，3回振った平均値であれば，3回連続1や6がでて，標本平均が1や6になるということは十分あり得るだろうが，100回振ったときの標本平均が，1や6になることは考えにくい。つまり，標本平均のばらつきは測定回数とともに減少していくということである。

標本平均の標本分散

標本平均の母分散は $V(\bar{x}) = \dfrac{\sigma^2}{n}$ で表されるが，先程と同様

σ^2が母数であるので，知ることができない。よって，データの母分散σ^2の代用としてデータの標本分散$s^2(x)$を用いることによって，標本平均の標本分散を求める。

$$\hat{\sigma}^2(\bar{x}) = s^2(\bar{x}) = \frac{s^2(x)}{n}$$

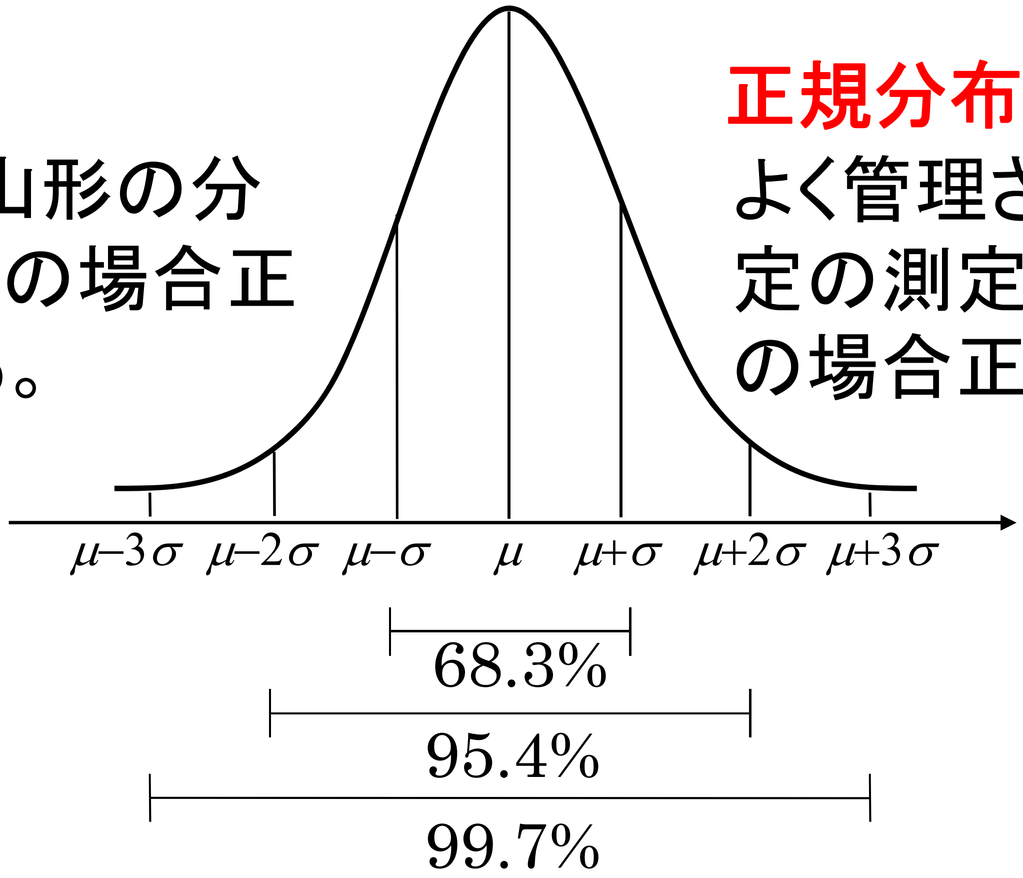
正規分布について

正規分布

よく見られる山形の分布はほとんどの場合正規分布である。

正規分布の性質

よく管理されている測定の測定値はほとんどの場合正規分布する。

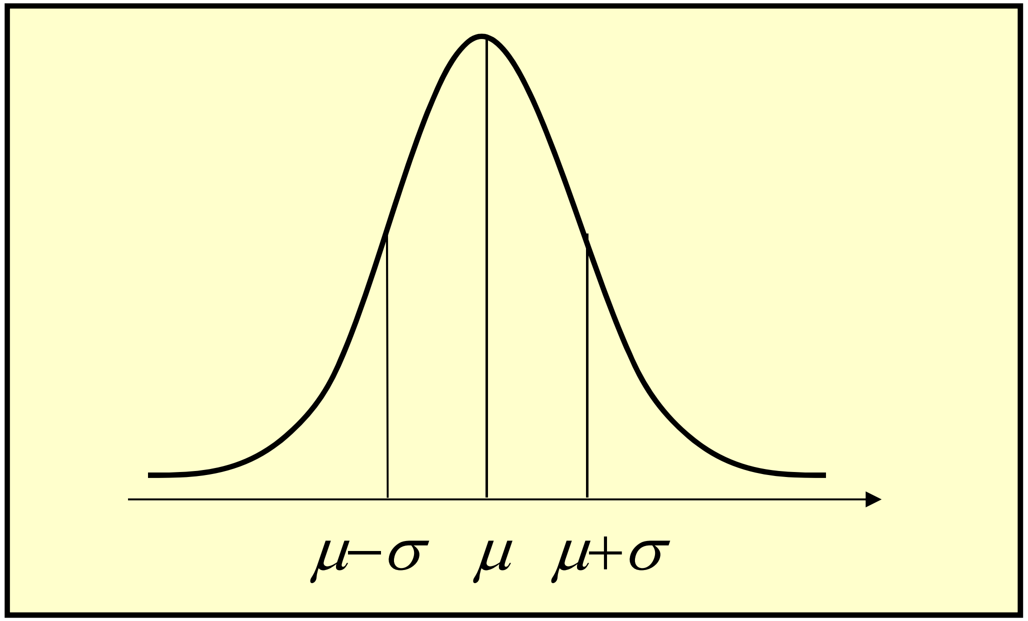


μ±1σ：68.3%
μ±2σ：95.4%
μ±3σ：99.7%

の値が含まれる。

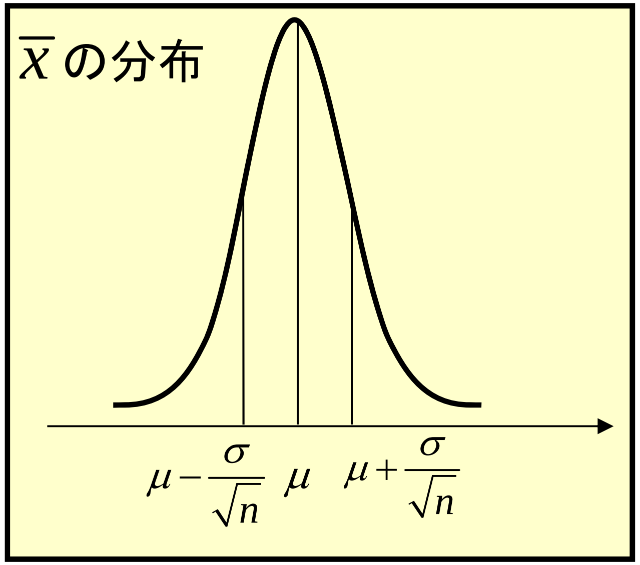
推定について

- 正規分布の性質を用いて推定を行う。
  ここで扱うのは母平均の推定である。
  測定の母集団は正規分布で，母平均がμ，母標準偏差がσであるとする。

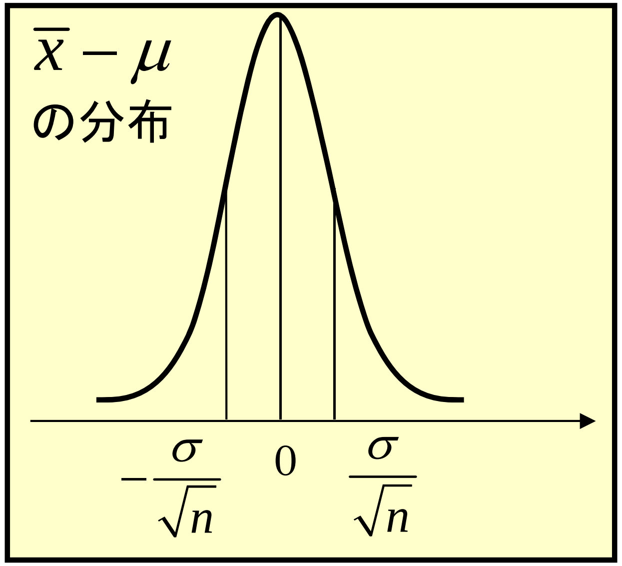


推定について

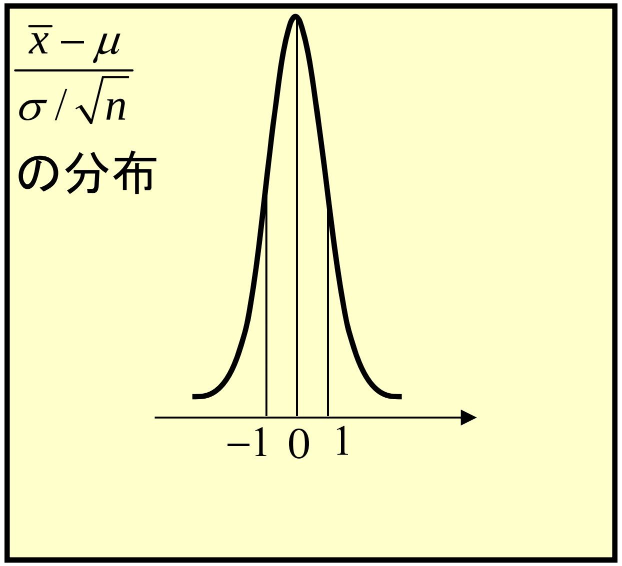
ここからデータをいくつかサンプリングして標本平均を算出



標準化を行う。まずは標本平均を0にする。



標準化を行う。標準偏差を1にする。



母平均μ, 母標準偏差σである正規分布をしている母集団からn個サンプリングを行い算出した標本平均を$\overline{x}$とするとき,

$\frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$ の分布は必ず母平均0, 母標準偏差1の正規分布をする

推定について

よって，

$$-1 < \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < 1$$

の区間には全データの68.3%が含まれる

$$-2 < \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < 2$$

の区間には全データの95.4%が含まれる

$$-3 < \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < 3$$

の区間には全データの99.7%が含まれる

推定について

代表で，$-2 < \dfrac{\overline{x} - \mu}{\sigma / \sqrt{n}} < 2$ を考える。

$-2 < \dfrac{\overline{x} - \mu}{\sigma / \sqrt{n}} < 2$ を変形する。

$$-2\frac{\sigma}{\sqrt{n}} < \overline{x} - \mu < 2\frac{\sigma}{\sqrt{n}}$$

$$-\overline{x} - 2\frac{\sigma}{\sqrt{n}} < -\mu < -\overline{x} + 2\frac{\sigma}{\sqrt{n}}$$

$$\overline{x} - 2\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + 2\frac{\sigma}{\sqrt{n}}$$

母平均の存在する区間を推定することができた。

推定について

- 任意の確率で範囲を推定するには‥正規分布表を用いる。

両側危険率	$\pm z$
20%	1.28
10%	1.65
5%	1.96
2%	2.33
1%	2.58
0.2%	3.29
0.02%	3.72

$$\bar{x} - z\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z\frac{\sigma}{\sqrt{n}}$$

推定の例

- 例：ある製品の濃度測定

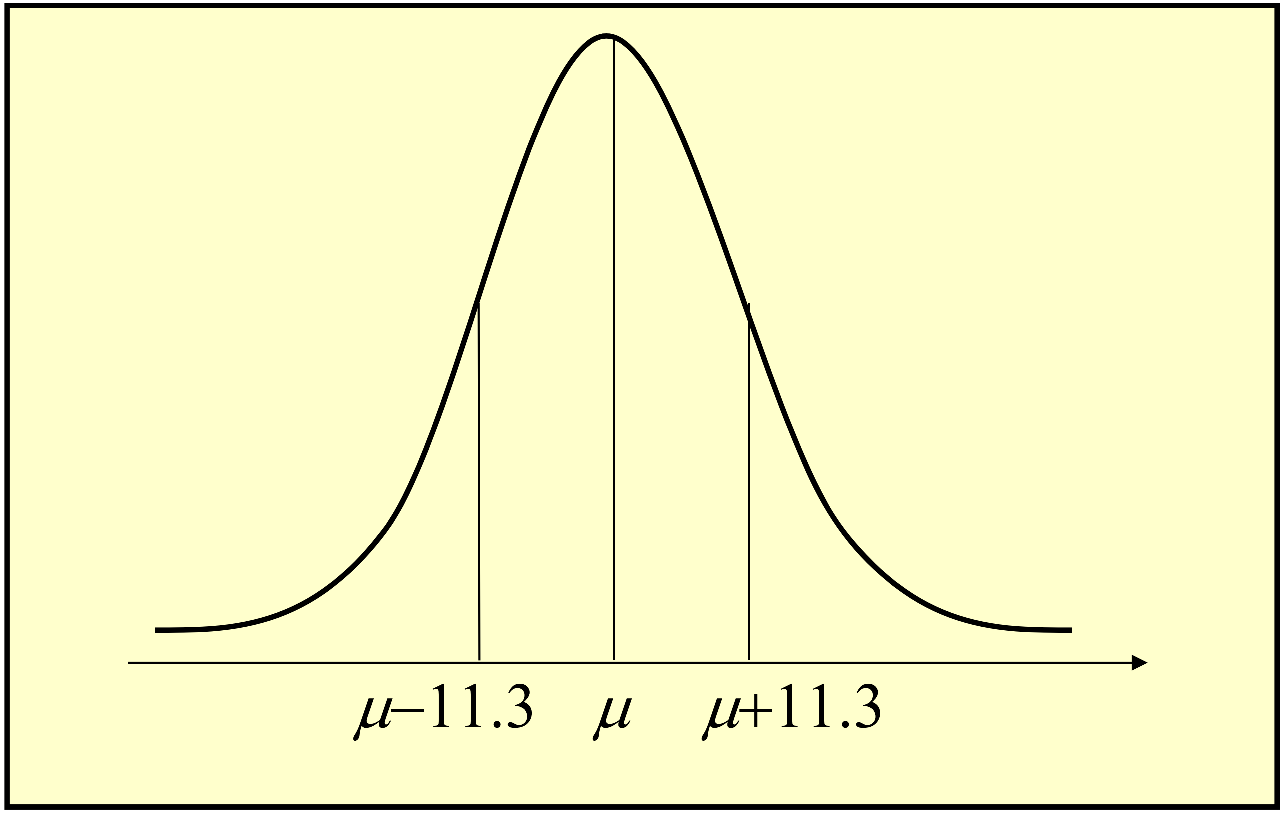
  ある溶液が生産ラインで制作され，ボトル詰めされている。その溶液には塩化物イオンが含まれている。そのボトル間での塩化物イオン濃度のばらつきは標準偏差が11.3 ppmの正規分布していることが分かっている。この製品群からボトル10個サンプリングを行い，濃度測定した結果

  標本平均 $\overline{x} = 126.2 \text{ ppm}$

  が得られた。この溶液の濃度の母平均が存在する範囲を信頼水準99%で推定せよ。

推定の例

- 質量の分布の図



推定の例

- 区間を推定

  99%の信頼水準ということは，両側危険率が1%のとき。よって，zの値は2.58

$$\bar{x} - z\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z\frac{\sigma}{\sqrt{n}}$$

$$126.2 - 2.58\frac{11.3}{\sqrt{10}} < \mu < 126.2 + 2.58\frac{11.3}{\sqrt{10}}$$

$$126.2 - 9.2 < \mu < 126.2 + 9.2$$

$$117.0 < \mu < 135.4$$

この製品の濃度の母平均は，117.0 ppmから135.4 ppmの間にあるということが，99%の信頼水準で分かった