代表的な確率分布

- 正規分布（ガウス分布）normal distribution, Gaussian distribution

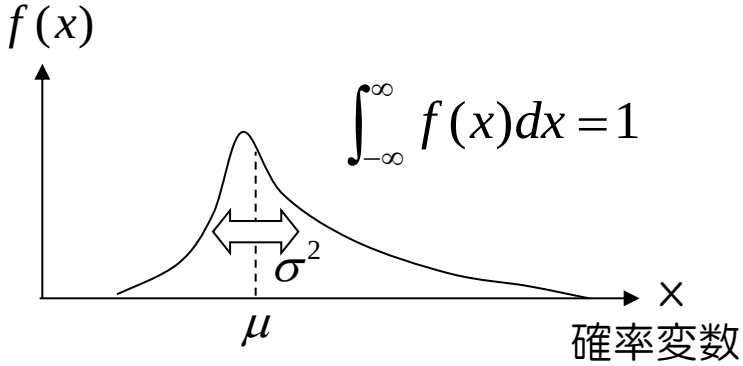
- 二項分布 binomial distribution
- ポアソン分布 Poisson distribution

- t-分布 (Student's t-distribution)

- χ^2分布(カイ２乗，カイスクエアと読む） χ^2 distribution

確率変数，確率密度関数

確率密度関数

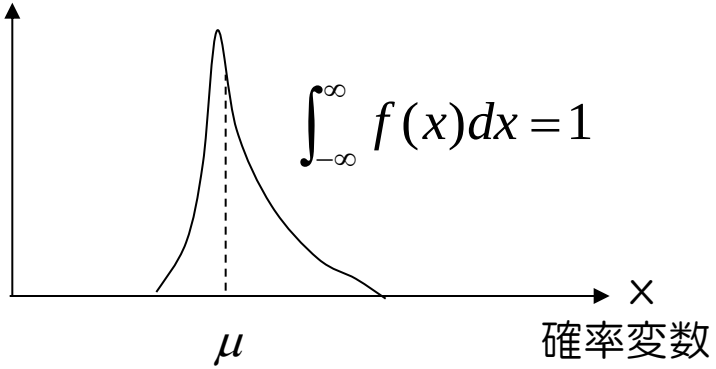


平均： $\mu = \int_{-\infty}^{\infty} x \cdot f(x)dx$

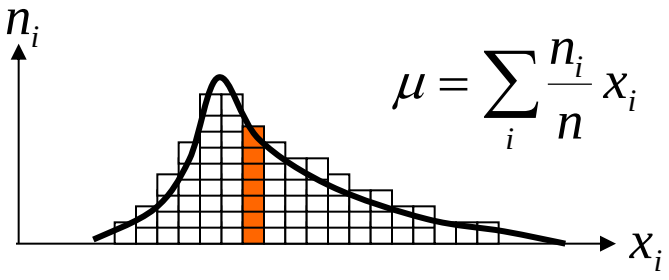
分散： $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)dx$

例）ある場所，ある日時での気温の確率．
x：気温
f（x）：気温xが起こる確率

$f(x)$ もし平均が同じで分散が小さいなら



度数　標本平均とのアナロジー



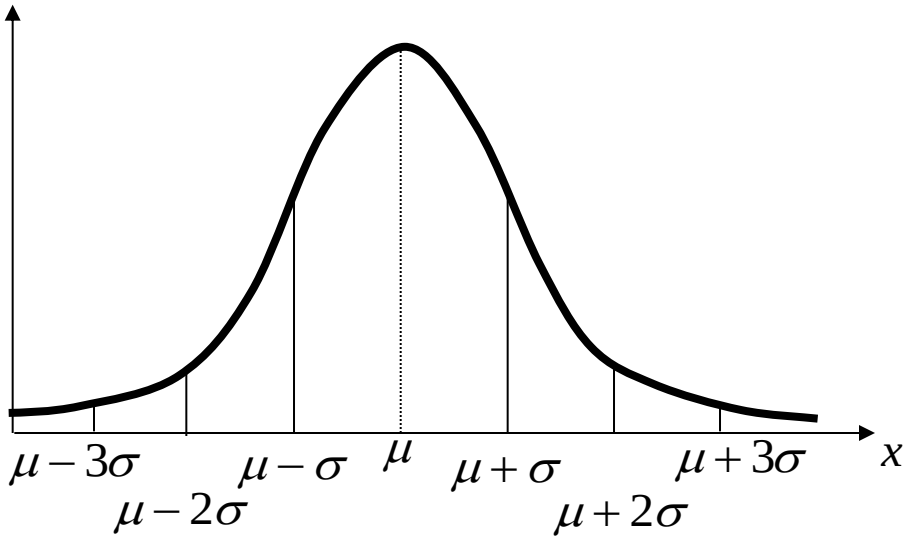
正規分布(ガウス分布)

確率密度関数

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

この分布の平均と分散は,

$$mean=\int_{-\infty}^{\infty}xf(x)dx=\int_{-\infty}^{\infty}x\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]dx=\mu$$



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

正規分布は平均μと分散σ^2によって
完全に記述される.

$N(\mu,\sigma^2)$ と表記する

確率変数の範囲と確率（よく用いられる値）

$\mu-\sigma\le x\le\mu+\sigma$ ----- 68.27%

$\mu-2\sigma\le x\le\mu+2\sigma$----- 95.45%

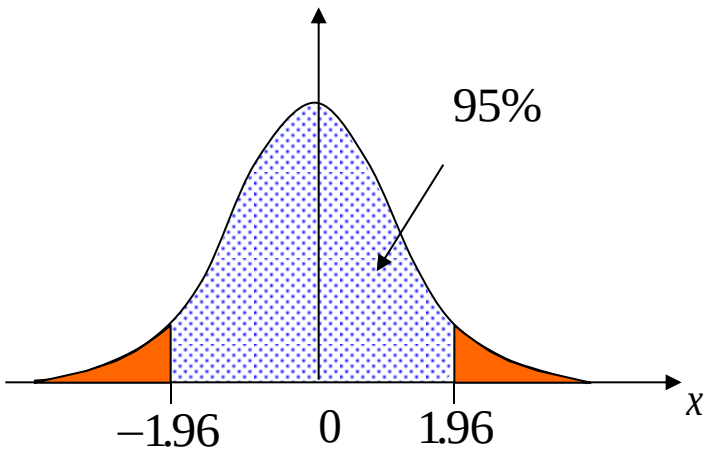
$\mu-3\sigma\le x\le\mu+3\sigma$ ----- 99.73%

$\mu-1.96\sigma\le x\le\mu+1.96\sigma$----- 95%

特に，平均０，分散１の正規分布
N(0,1)を標準正規分布と呼ぶ.

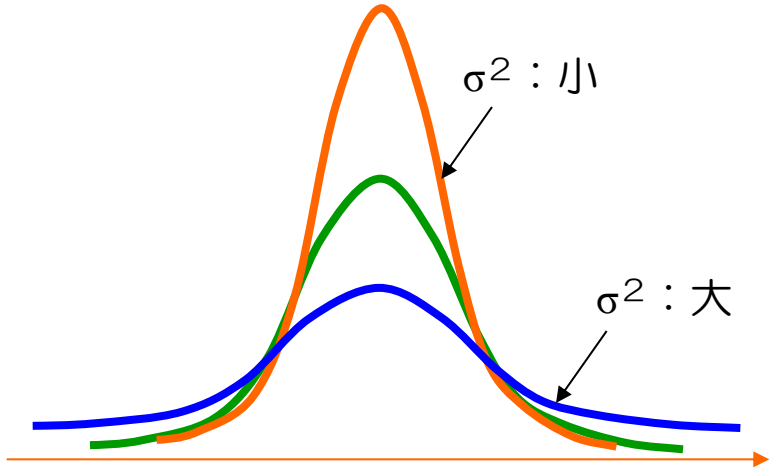
正規分布（ガウス分布）つづき $f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$

標準正規分布 $N(0,1)$

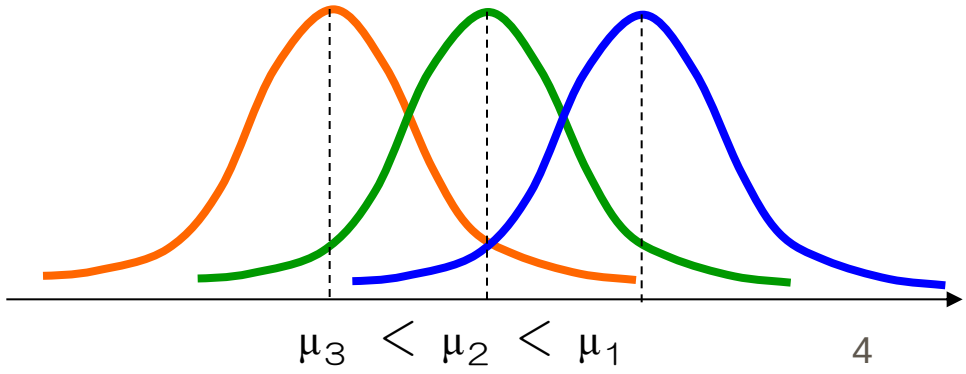


９５％の確率で存在する範囲が
統計ではしばしば使われる.
標準正規分布では－１．９６から
１．９６の範囲となる.

平均が同じで分散が異なる正規分布



分散が同じで平均が異なる正規分布



二項分布 binomial distribution

例）３回サイコロを投げて，x回，１の目が出る確率を考える．

	０回	１回	２回	３回
$P(x)$	$\left(\frac{5}{6}\right)^3$	$3\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^2$	$3\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^1$	$\left(\frac{1}{6}\right)^3$
	$\neq 1$ $\neq 1$ $\neq 1$	1 $\neq 1$ $\neq 1$	1 1 $\neq 1$	1 1 1

$\Rightarrow P(x)={}_3C_x\left(\frac{1}{6}\right)^x\left(\frac{5}{6}\right)^{3-x}$

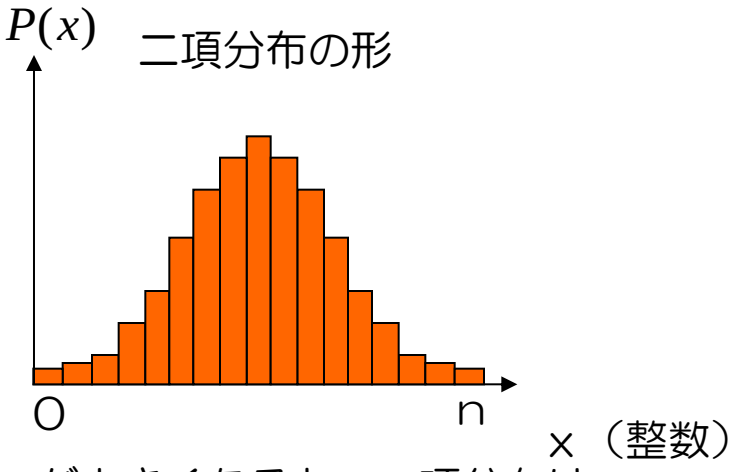
一般に，確率ｐをもつ事象が，
ｎ回の観察でｘ回起こる確率P(x)は

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

この式で表される確率分布を二項分布と呼ぶ．

平均：$\mu=np$

分散：$\sigma^2=np(1-p)$



ｎが大きくなると，二項分布は
正規分布に近づく

ポアソン分布 Poisson distribution

二項分布において，実験回数ｎが十分大きい場合，
二項分布はポアソン分布で近似できる．

$$P(x)={}_nC_x p^x (1-p)^{n-x}$$

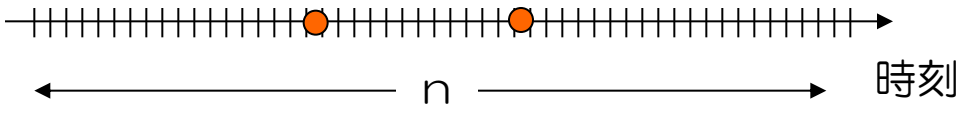
⇩ 近似

$$P(x)=\frac{\mu^x e^{-\mu}}{x!} \qquad \text{ただし} \quad \mu = np$$

平均μが大きければ，ポアソン分布は
正規分布に近似できる．

例）千葉市の１日あたりの交通事故件数の確率分布

１日を十分細かくきざんで考える（例えば１分単位）．
すると，このきざみのなかでは，事故が起こるか起こらないかの，どちらかの事象のみ起こるとみなせる．
１つのきざみ内で事故が起こる確率をｐとすれば，
１日にｘ件事故が起こる確率は，二項分布で表せる．



１日平均５回，事故が起こるとする．

１）二項分布で考えると，

１分あたりに事故が起こる確率は

$$p = 5/(24\times 60)$$

ある１日に，ｘ回起こる確率は，

$$P(x)={}_{24\times 60}C_x p^x (1-p)^{24\times 60-x}$$

２）ポアソン分布で考えると

$$P(x)=\frac{5^x e^{-5}}{x!}$$

事故数	二項分布	ポアソン分布
0	0.00668	0.00674
1	0.03351	0.03369
2	0.08402	0.08422
3	0.14032	0.14037
4	0.17565	0.17547
5	0.17577	0.17547
6	0.14648	0.14622
7	0.10455	0.10444
8	0.06526	0.06528
9	0.03618	0.03627
10	0.01804	0.01813

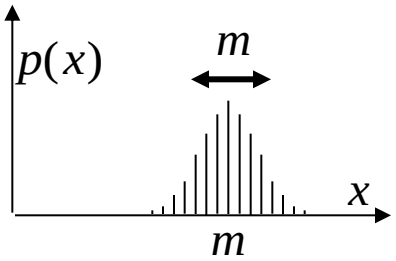
← 5回

ポアソン分布の性質とフォトンノイズの例

ポアソン分布は，平均と分散が等しい．

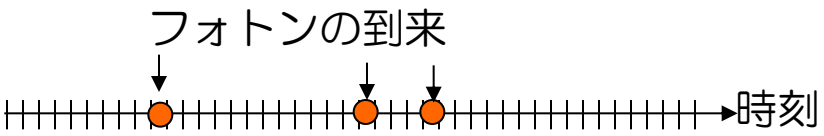
$$P(x)=\frac{m^x e^{-m}}{x!}$$

において　平均＝分散＝m



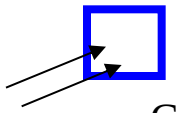
標準偏差は　$\sigma=\sqrt{m}$

例）明るい条件と暗い条件で，単位時間あたりにCCDの画素に到達するフォトン数を考える．



CCDの画素に到達するフォトン数はポアソン分布に従う．

[暗い]



CCD画素

平均をm=100とする

[明るい]

CCD画素

平均をm=10000とする

標準偏差は

$\sigma=\sqrt{100}=10$　　$\sigma=\sqrt{10000}=100$

フォトン数ｘのちらばりを
±２σの範囲で考えると

$80<x<120$　　$9800<x<10200$

カメラのゲインコントロールによって明るさを合わせられることを考えて，それぞれの平均が１００になるように正規化すると

$80<x<120$　　$98<x<102$

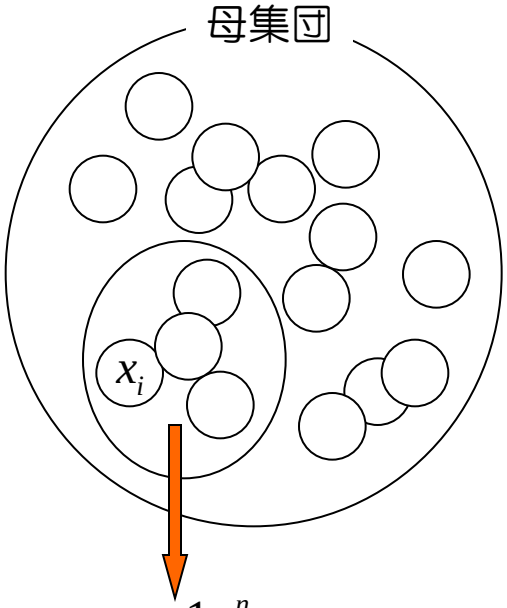
以上より，暗い状態ではノイズが増えることがわかる（フォトンノイズという）

中心極限定理 central limit theorem

分布がどのようなものであっても，平均値μ，分散σ^2を
もつ母集団からとられた大きさｎの標本の平均値の分布は，
ｎが大きくなるとき，正規分布$N(\mu, \sigma^2/n)$に近づく．
したがって，

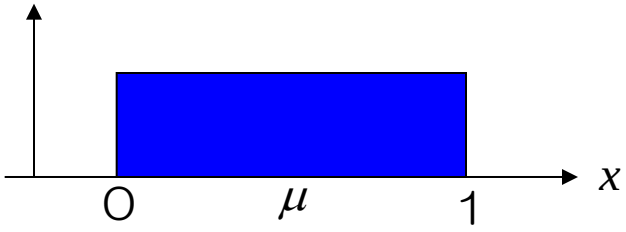
$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

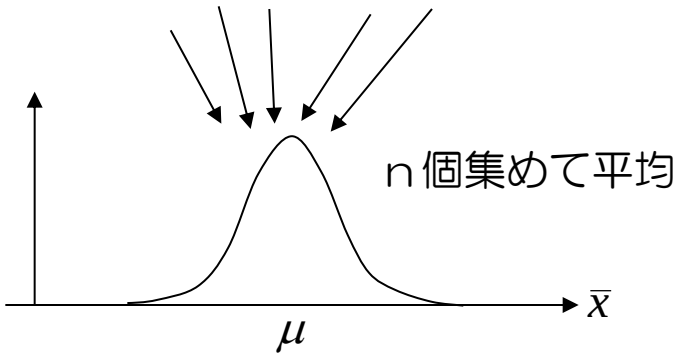
の分布は，ｎが大となるとき，標準正規分布に近づく．

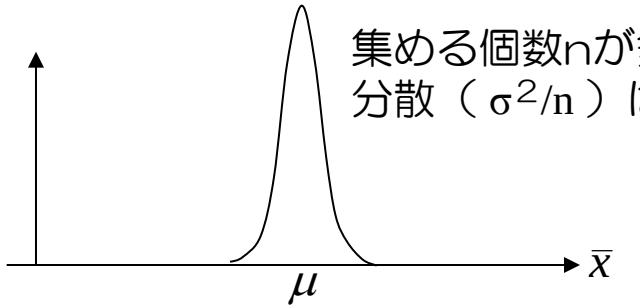


$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

例）母集団の分布が一様分布の場合







集める個数nが多いほど
分散（σ^2/n）は小さい．

中心極限定理：多くの観測値を正規分布で近似する裏付けとなっている

サンプルから母集団統計量を推定する

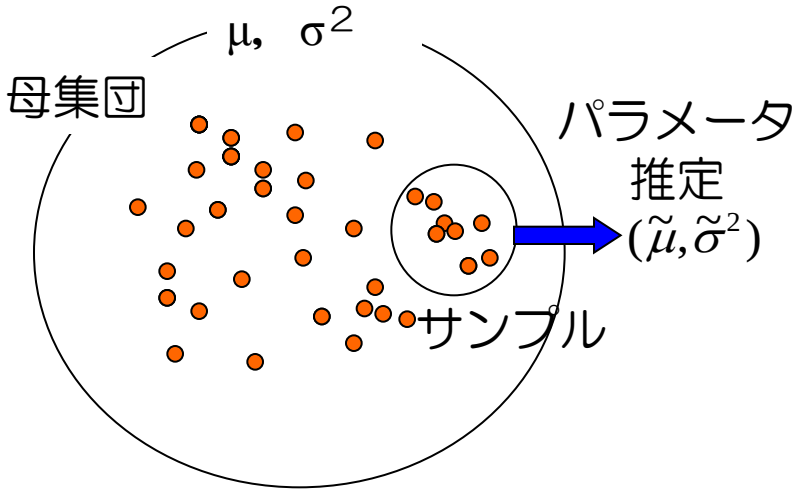
命題：
得られたサンプルから，
その発生母体である母集団の統計量を推定したい．

平均：１次の統計量

分散：２次の統計量

例）母集団が正規分布の場合

母集団を表すパラメータは平均μと分散σ^2のふたつである．



平均：　$\mu = \int_{-\infty}^{\infty} x f(x) dx$　（１次）

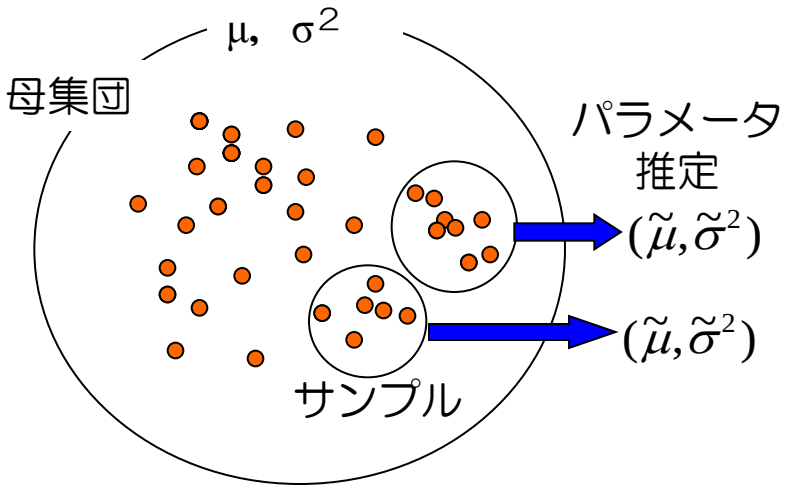
分散：　$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$　（２次）

不偏推定量　unbiased estimator
―平均の不偏推定量―

不偏推定量とは，サンプルから求めた
母集団パラメータの期待値が，真の
母集団パラメータに一致するものをいう．

例）母集団が正規分布の場合

母集団を表すパラメータは平均μと分散σ^2
のふたつである．



$$E\{\tilde{\mu}\}=\mu ?$$
$$E\{\tilde{\sigma}^2\}=\sigma^2 ?$$

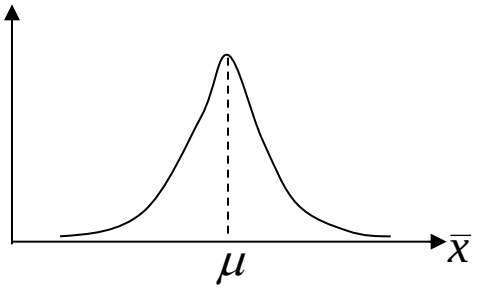
母集団平均の推定をサンプル平均で行った
場合，

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

サンプル平均の期待値は

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

となり，母集団平均に一致する．
よって，サンプル平均は，母集団
平均に対する不偏推定量といえる．



分散の不偏推定量

標本分散の期待値を計算してみる

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

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

$$= \frac{1}{n}E\{\sum_{i=1}^{n}(x_i - \mu)^2\}$$

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

$$+\frac{1}{n}E\{n(\bar{x} - \mu)^2\}$$

上式右辺の第１項は

$$\frac{1}{n}E\{\sum_{i=1}^{n}(x_i - \mu)^2\} = \frac{1}{n}\sum_{i=1}^{n}E\{(x_i - \mu)^2\}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sigma^2 = \sigma^2$$

第３項は

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

$$= E\{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)\right)^2\}$$

$$= E\{\frac{1}{n^2}\sum_{i}\sum_{j}(x_i - \mu)(x_j - \mu)\}$$

$$= E\{\frac{1}{n^2}\sum_{i}(x_i - \mu)^2\}$$

$$= \frac{1}{n^2}\sum_{i}E\{(x_i - \mu)^2\}$$

$$= \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

第２項も同様に計算できる．結局，

$$E\{s^2\} = \sigma^2 - \frac{2}{n}\sigma^2 + \frac{1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

となり，母集団分散には一致しないことがわかる

n-1で割れば母集団分散に一致することを確認しなさい．

分散の不偏推定量（つづき）直感的解釈

なぜ分散の推定を，（ｎで割らずに）

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

で与えるか？　⇒直感的解釈

仮に母集団の平均μが既知であれば，
ｎ個のデータからの分散の推定は

$$\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu\right)^2$$

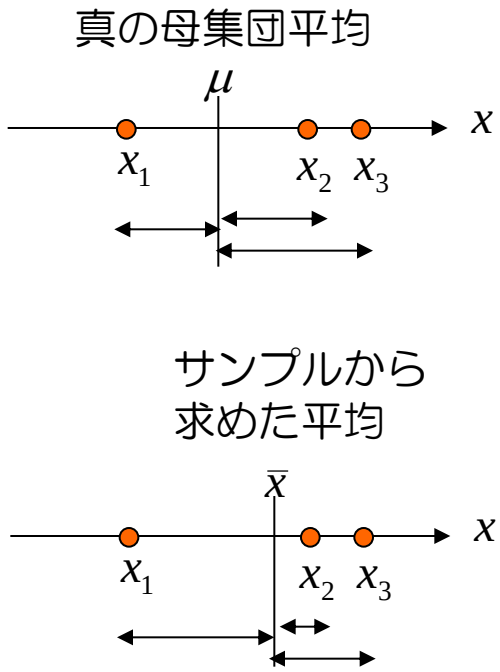
で与えればよい．これに対し，母集団平均μが未知のために，かわりにサンプル平均を用いた場合の分散をｓ^{2}とすると，

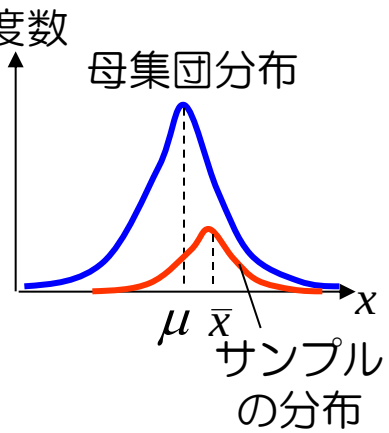
$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

この場合，かならず

$$s^2 \le \bar{\sigma}^2$$

が成り立つ．すなわち，ｓ^{2}は真の
母集団分散を過小に推定する傾向がある．
そこで，ｎで割らずにｎ－１で割ることで
この過小推定を防ぐ．





サンプルから母集団の平均を推定する

母集団が正規分布に従うとする.
もし，母集団正規分布の平均と分散が既知なら
ｎ個のサンプルを集めてきて得た平均値は
ｎの値によらず，正規分布$N(\mu,\sigma^2/n)$をする.

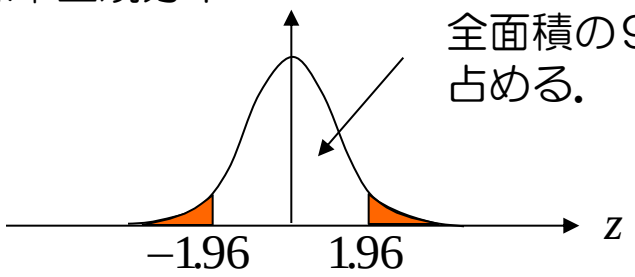
⇕

標準化（平均を引き，標準偏差で割る）を
行えば，その値は標準正規分布に従う.

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

いま，母集団の分散σ^2のみが既知としたとき,
標本から推定される母集団平均μの区間を考える.

標準正規分布：



標準正規分布は－１．９６から１．９６の
間をとる確率が９５％である.

$$P\left(-1.96 < \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < 1.96\right) = 0.95$$

カッコの中を書き直せば,

$$P\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

これより，未知の母集団平均μが

$$\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

という範囲に９５％の確率で存在する
ことがわかる.