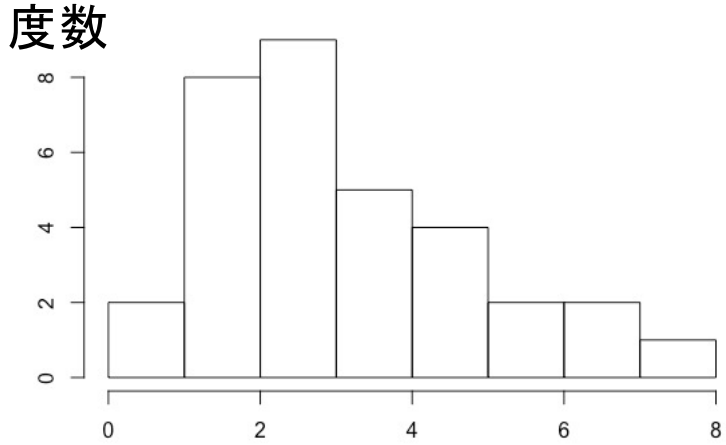
多変量解析

第3回　データの集約

萩原•篠田
情報理工学部

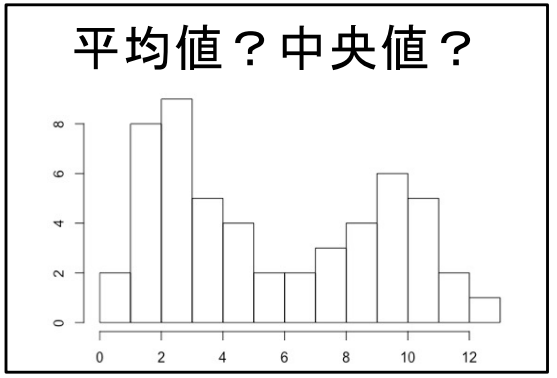
代表値・散布度

データ									
1	1	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	4
4	4	4	4	5	5	5	5	6	6
7	7	8							



代表値（中心的傾向）

平均値（mean）　　: 3.606
中央値（median）　: 3
最頻値（mode）　　: 3

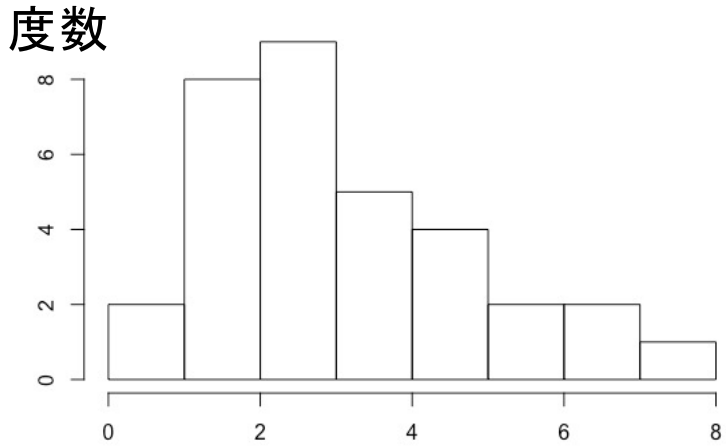


散布度（分布、散らばり具合）

分散（偏差の2乗の平均）　　: 3.12
標準偏差（分散の平方根）　　: 1.77（= $\sqrt{3.12}$）
範囲（最大−最小値）　　: 7（= 8 - 1）
四分位範囲（上位−下位四分位数）: 3（= 5 - 2）

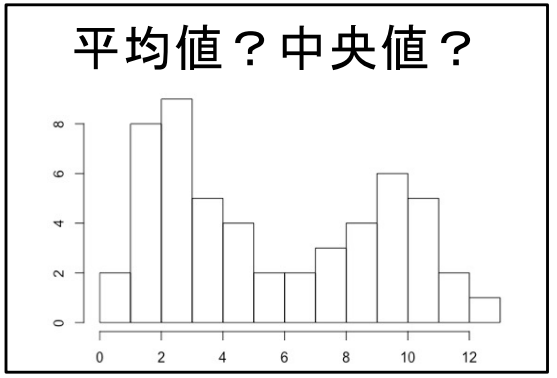
代表値・散布度

データ									
1	1	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	4
4	4	4	4	5	5	5	5	6	6
7	7	8							



代表値（中心的傾向）

平均値（mean）　　: 3.606
中央値（median）　　: 3
最頻値（mode）　　: 3

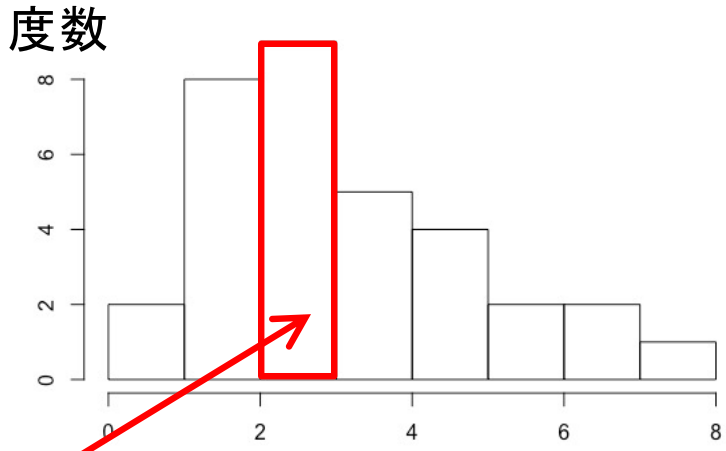


散布度（分布、散らばり具合）

分散（偏差の2乗の平均）　　: 3.12
標準偏差（分散の平方根）　　: 1.77（= $\sqrt{3.12}$）
範囲（最大−最小値）　　: 7（= 8 - 1）
四分位範囲（上位−下位四分位数）: 3（= 5 - 2）

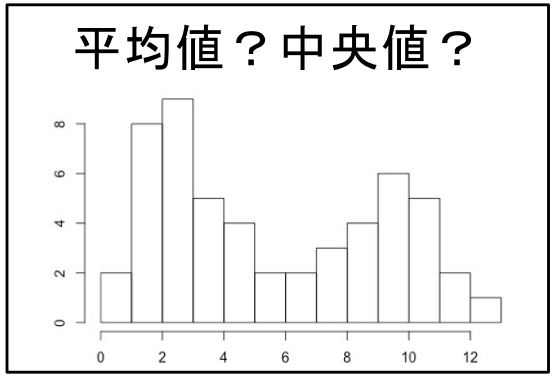
代表値・散布度

データ									
1	1	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	4
4	4	4	4	5	5	5	5	6	6
7	7	8							



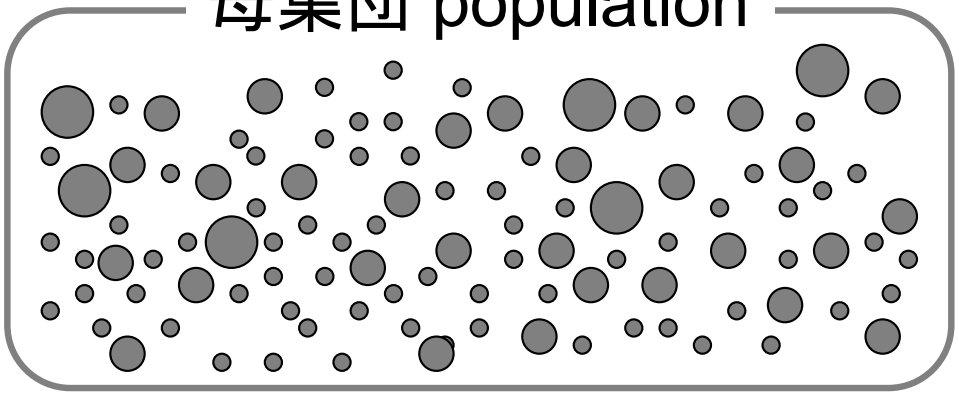
代表値（中心的傾向）

平均値（mean）　　：3.606
中央値（median）　：3
最頻値（mode）　　：3



散布度（分布、散らばり具合）

分散（偏差の2乗の平均）　　：3.12
標準偏差（分散の平方根）　　：1.77（$= \sqrt{3.12}$）
範囲（最大−最小値）　　：7（= 8 - 1）
四分位範囲（上位−下位四分位数）：3（= 5 - 2）



平均値 mean, average

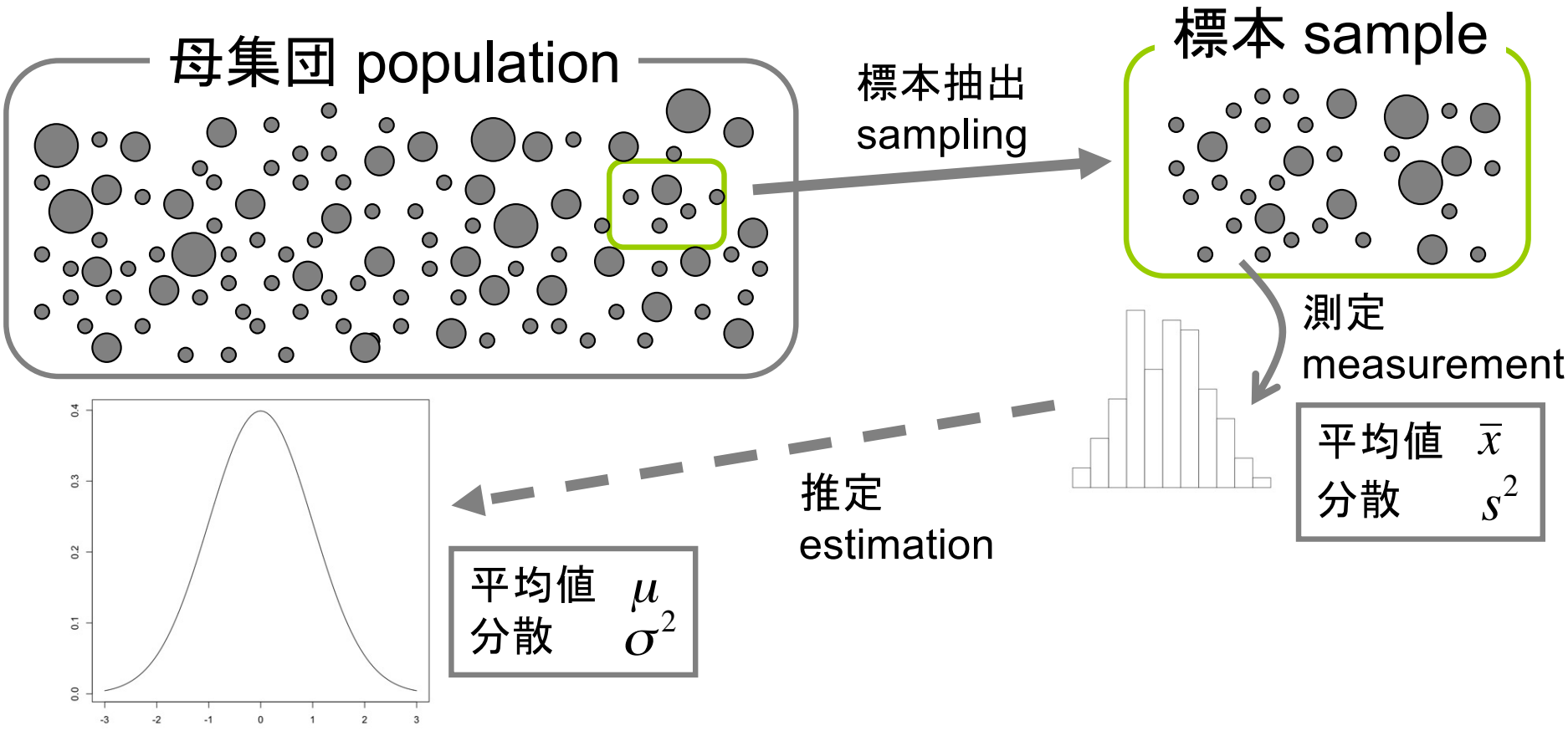
$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}(x_1 + x_2 + x_3 + \cdots + x_N)$$

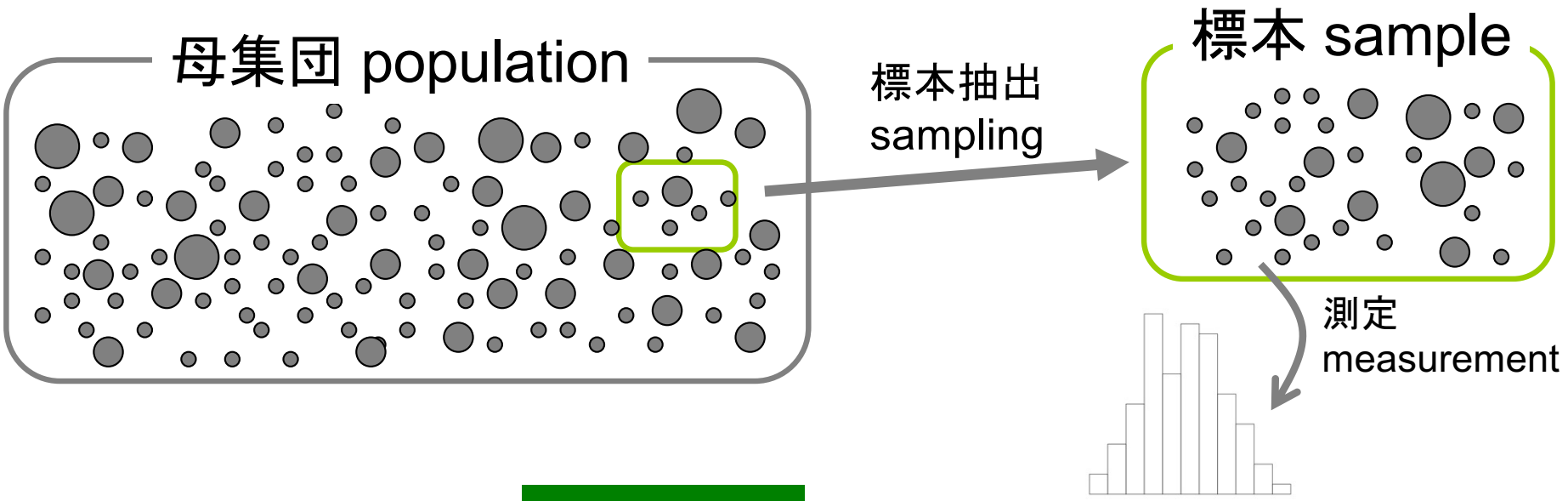
分散 variance

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

標準偏差 standard deviation, SD

$$\sigma$$


母集団 population
標本抽出
sampling
標本 sample
測定
measurement
平均値 $\bar{x}$
分散 s^2
推定
estimation
平均値 μ
分散 σ^2



平均値 mean, average

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

期待値

不偏推定値

平均値

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

分散 variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 > E\{s_n{}^2\}$$

過小評価

分散

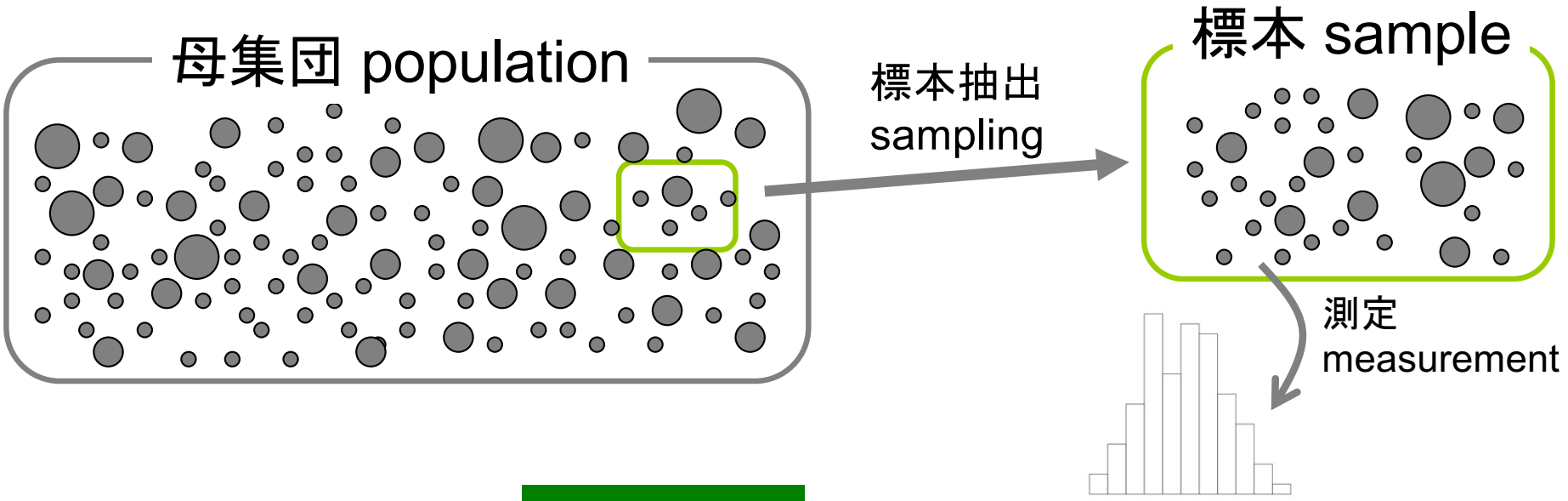
$$s_n{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

標準偏差 standard deviation, SD

$$\sigma > E\{s_n\}$$

標準偏差

$$s_n$$



平均値 mean, average

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

期待値

不偏推定値

平均値

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

分散 variance

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

分散

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

不偏推定値

標準偏差 standard deviation, SD

$$\sigma = E\{s\}$$

標準偏差

$$s$$

なぜ分散の不偏推定値は

$$s_n^{\,2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{ではなく,} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{なのか？}$$

母集団の分散

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

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

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

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

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

$$= E\left\{s_n^{\,2}\right\} + \frac{\sigma^2}{n}$$

$$\sum_{i=1}^{n}\frac{(x_i - \bar{x})}{n} = \sum_{i=1}^{n}\frac{x_i}{n} - \sum_{i=1}^{n}\frac{\bar{x}}{n} = \bar{x} - \bar{x} = 0$$

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

$$\sigma^2 = \frac{n}{n-1}E\left\{s_n^{\,2}\right\}$$

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

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

$$= E\left\{s^2\right\}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \quad n\text{個が独立}$$

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

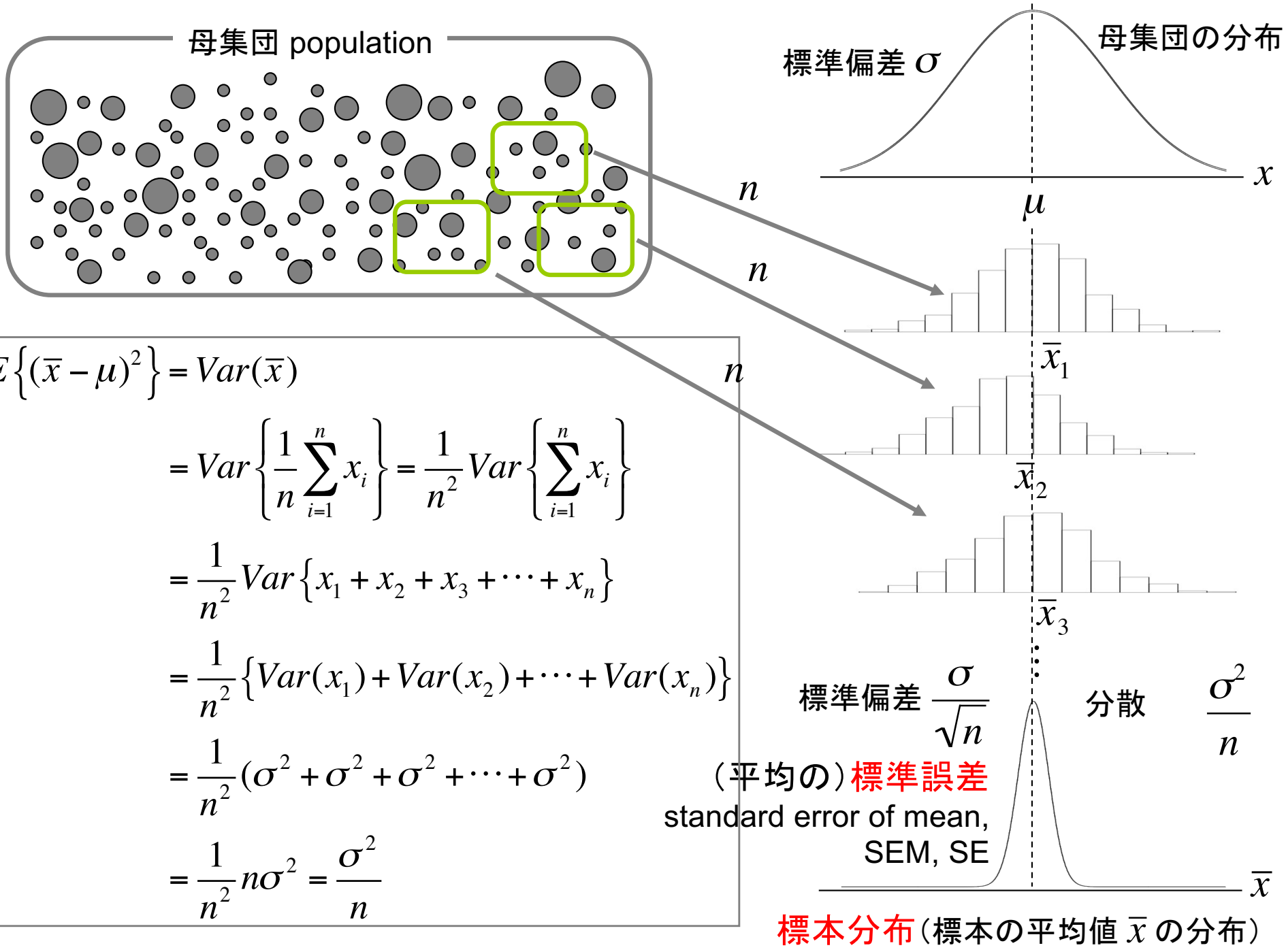
$$\bar{x} = \sum_{i=1}^{n}\frac{x_i}{n}$$

$\bar{x}$ の関係式があるため
(n-1)個の偏差が決まれば
残り1個の偏差が決まる
つまり自由度はnでなくn-1

標本分布（標本の平均値 $\bar{x}$ の分布）の分散

（参考）標本分布の標準偏差 = 標準誤差（SE、SEM）

標本分布（標本の平均値 $\bar{x}$ の分布）と標準誤差（標本分布の標準偏差）



$$E\left\{(\bar{x}-\mu)^2\right\} = Var(\bar{x})$$

$$= Var\left\{\frac{1}{n}\sum_{i=1}^{n} x_i\right\} = \frac{1}{n^2} Var\left\{\sum_{i=1}^{n} x_i\right\}$$

$$= \frac{1}{n^2} Var\left\{x_1 + x_2 + x_3 + \cdots + x_n\right\}$$

$$= \frac{1}{n^2}\left\{Var(x_1) + Var(x_2) + \cdots + Var(x_n)\right\}$$

$$= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \sigma^2 + \cdots + \sigma^2)$$

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

独立変数の結合分布

期待値（平均値）

$$E\{X \pm Y\} = E\{X\} \pm E\{Y\}$$

$$E\{aX\} = aE\{X\}$$

分散（の期待値）

$$\begin{aligned}
Var\{X \pm Y\} &= E\left[\{(X \pm Y) - (\bar{X} \pm \bar{Y})\}^2\right] \\
&= E\left[\{(X - \bar{X}) \pm (Y - \bar{Y})\}^2\right] \\
&= E\left[(X - \bar{X})^2 \pm 2(X - \bar{X})(Y - \bar{Y}) + (Y - \bar{Y})^2\right] \\
&= E\{(X - \bar{X})^2\} \pm 2E\{(X - \bar{X})(Y - \bar{Y})\} + E\{(Y - \bar{Y})^2\} \\
&= E\{(X - \bar{X})^2\} + E\{(Y - \bar{Y})^2\} \\
&= Var\{X\} + Var\{Y\}
\end{aligned}$$

$$\begin{aligned}
Var\{aX\} &= E\{(aX - a\bar{X})^2\} \\
&= E\{a^2(X - \bar{X})^2\} \\
&= a^2 E\{(X - \bar{X})^2\} \\
&= a^2 Var\{X\}
\end{aligned}$$

共分散

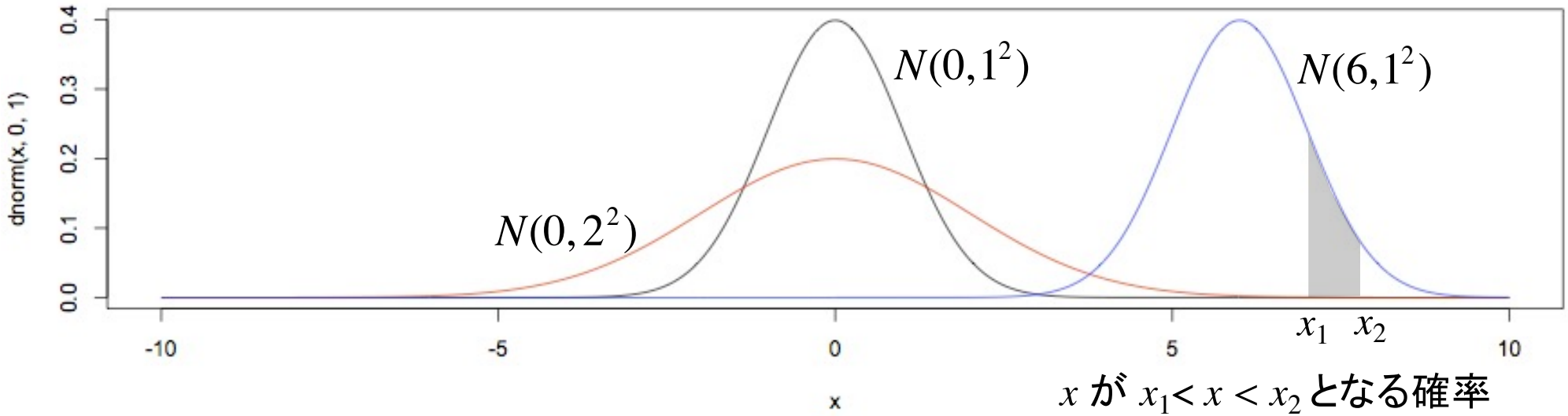
$$\begin{aligned}
E\{(X - \bar{X})(Y - \bar{Y})\} &= E\{XY - \bar{X}Y - X\bar{Y} + \bar{X}\bar{Y}\} \\
&= E\{XY\} - \bar{X}E\{Y\} - E\{X\}\bar{Y} + \bar{X}\bar{Y} \\
&= E\{XY\} - \bar{X}\bar{Y} \\
&= 0 \quad (when\ X, Y\ are\ independent)
\end{aligned}$$

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

$$N(\mu, \sigma^2)$$

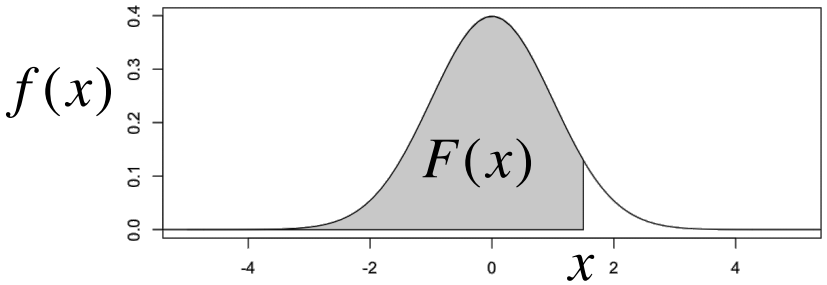
• 確率密度関数 $f(x)$
（probability density function, p.d.f.）

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



x が $x_1 < x < x_2$ となる確率

$$P(x_1 < x < x_2) = \int_{x_1}^{x_2} f(x)dx = \int_{-\infty}^{x_2} f(x)dx - \int_{-\infty}^{x_1} f(x)dx = F(x_2) - F(x_1)$$



• 累積分布関数 $F(x)$
（cumulative distribution function, c.d.f.）

$$P(x < x_1) = P(x \leq x_1) = F(x_1)$$

標準正規分布（standard normal distribution） $N(0,1)$

- 確率密度関数（p.d.f.）

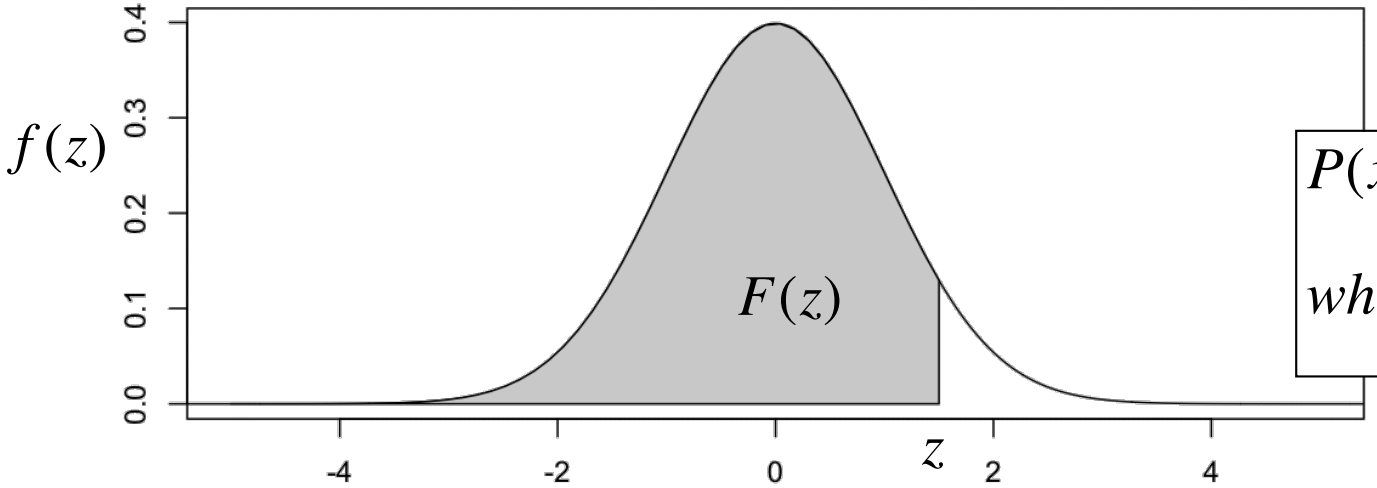
$$N(\mu,\sigma^2)$$

標準化

$$z=\frac{x-\mu}{\sigma}$$

$$N(0,1^2)$$

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



$$P(x<x_1)=P(z<z_1)$$

$$where \quad z=\frac{x-\mu}{\sigma}$$

- 累積分布関数（c.d.f.）

$$F(z)=\int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)dt$$

$$P(x<x_1)=F(x_1)=\int_{-\infty}^{x_1}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}dx=\int_{-\infty}^{z_1}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{z^2}{2}\right)\sigma\,dz=\int_{-\infty}^{z_1}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)dz$$

$$=F(z_1)=P(z<z_1)$$

$$F(z)=\int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)dt$$

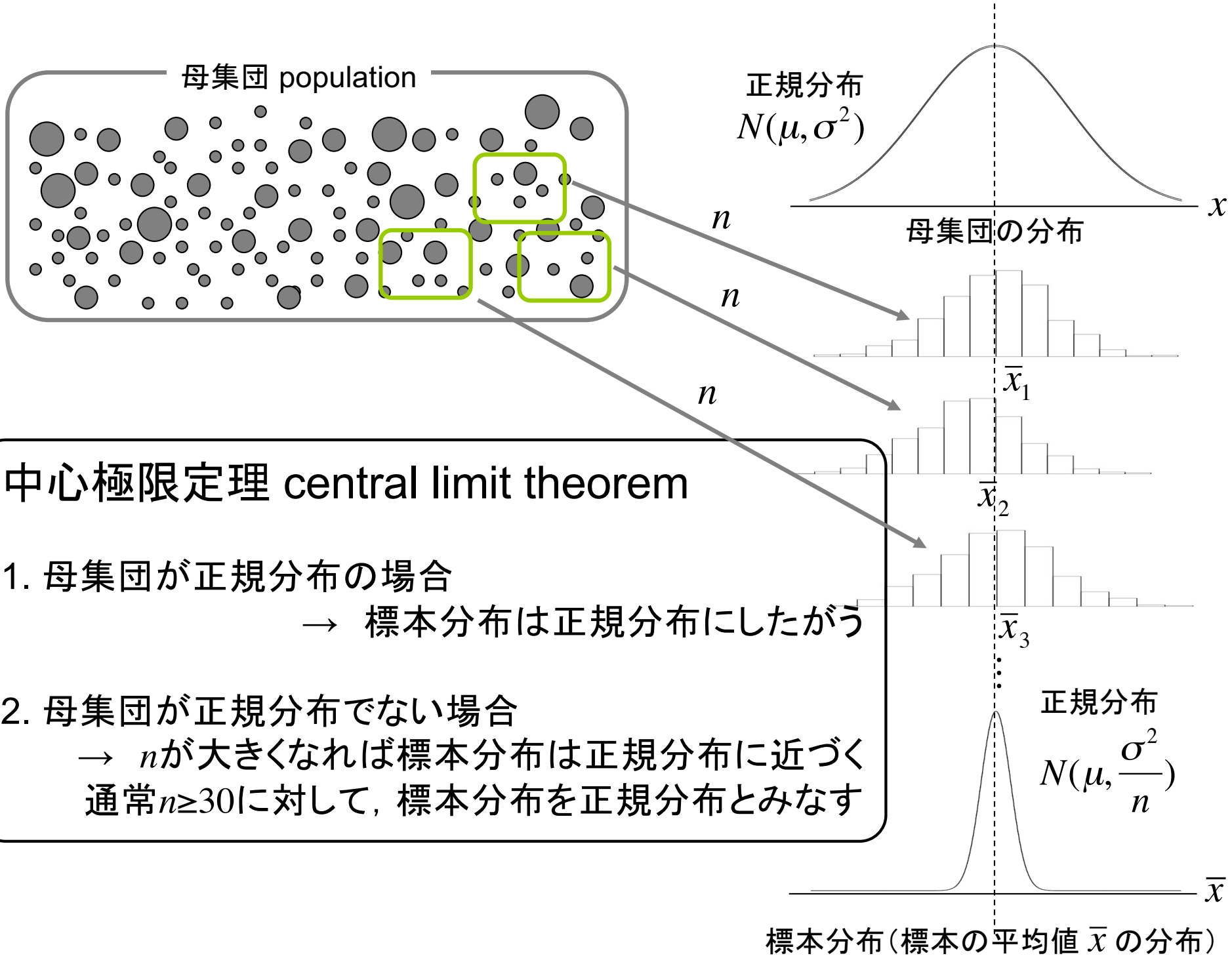
例：$N(\mu, \sigma^2)$で
$\mu-\sigma < x < \mu+\sigma$
となる確率は？
↓
$N(0, 1)$で
$-1 < z < 1$となる
確率に等しい

$P(-1< z <1)$
$=2(F(1)-F(0))$
$=2*(0.8413-0.5)$
$=0.6826$

問：$N(\mu, \sigma^2)$で
$\mu-2\sigma < x < \mu+2\sigma$
となる確率は？

表：$F(z)$標準正規分布$N(0, 1)$のc.d.f.

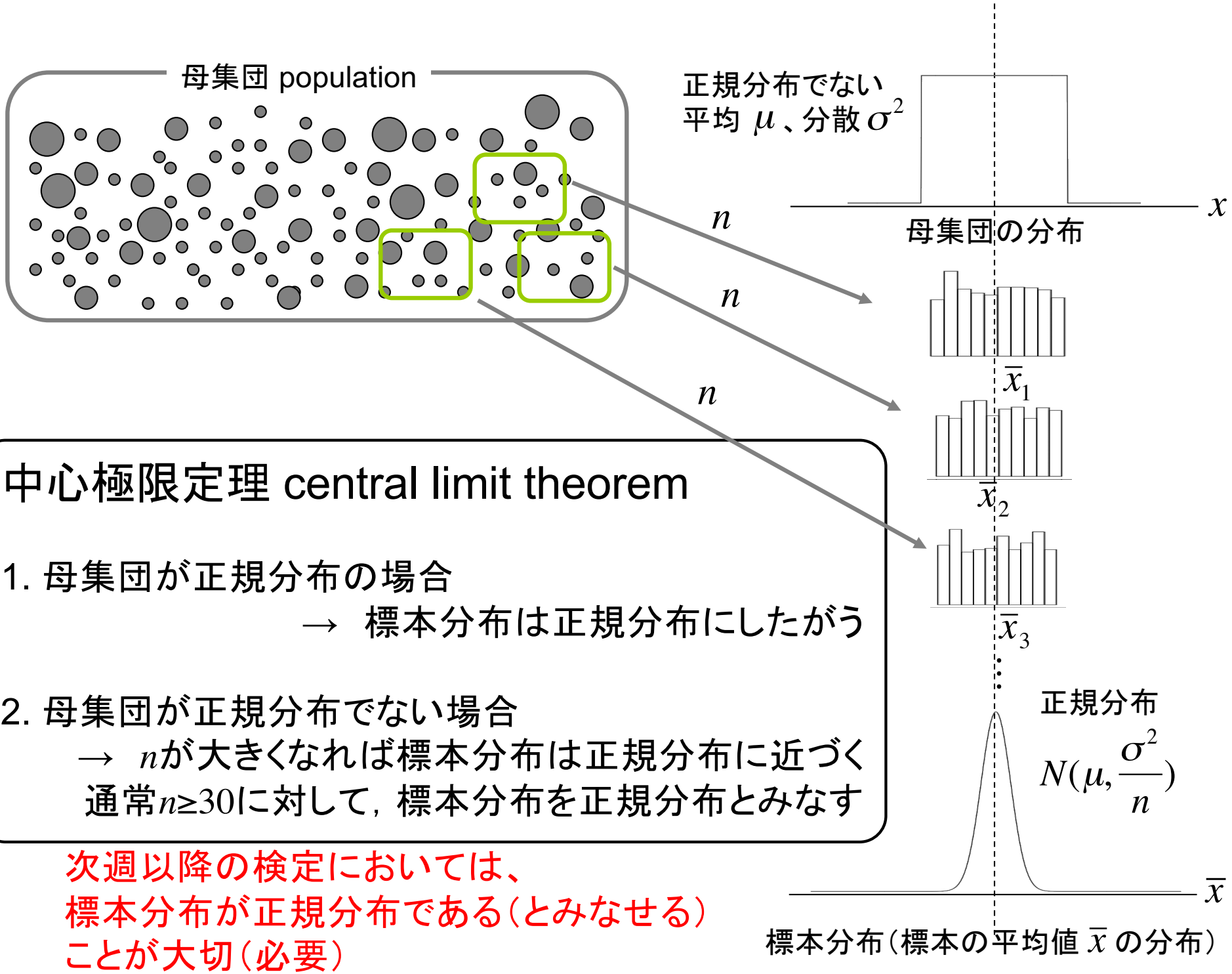
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990



中心極限定理 central limit theorem

1. 母集団が正規分布の場合
   → 標本分布は正規分布にしたがう

2. 母集団が正規分布でない場合
   → nが大きくなれば標本分布は正規分布に近づく
   通常$n \geq 30$に対して，標本分布を正規分布とみなす



中心極限定理 central limit theorem

1. 母集団が正規分布の場合
   → 標本分布は正規分布にしたがう

2. 母集団が正規分布でない場合
   → nが大きくなれば標本分布は正規分布に近づく
   通常$n \geq 30$に対して，標本分布を正規分布とみなす

次週以降の検定においては、
標本分布が正規分布である（とみなせる）
ことが大切（必要）

$$F(z)=\int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)dt$$

例：$N(\mu, \sigma^2)$で
$\mu-\sigma < x < \mu+\sigma$
となる確率は？
↓
$N(0, 1)$で
$-1 < z < 1$となる
確率に等しい

$P(-1< z <1)$
$=2(F(1)-F(0))$
$=2*(0.8413-0.5)$
$=0.6826$

問：$N(\mu, \sigma^2)$で
$\mu-2\sigma < x < \mu+2\sigma$
となる確率は？

表：$F(z)$標準正規分布$N(0, 1)$のc.d.f.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

（問題１）英国人女性の母集団の体重の分布は平均60kg，標準偏差5kgの正規分布にしたがう．この母集団から選んだ一人の女性の体重が70kg以上である確率を求めよ．

（問題２）ある機械は袋に詰める砂糖の重さが平均1000g，標準偏差5gの正規分布に従うように調整されている．袋を9個ずつ取り出して砂糖の重さの平均値を求め，これを繰り返すとき，平均値の分布はどうなるか．

（問題３）問題２の機械が詰めた袋を無作為に9個取り出して砂糖の重さの平均値を量るとき，その値が1003gより重くなる確率を求めよ．