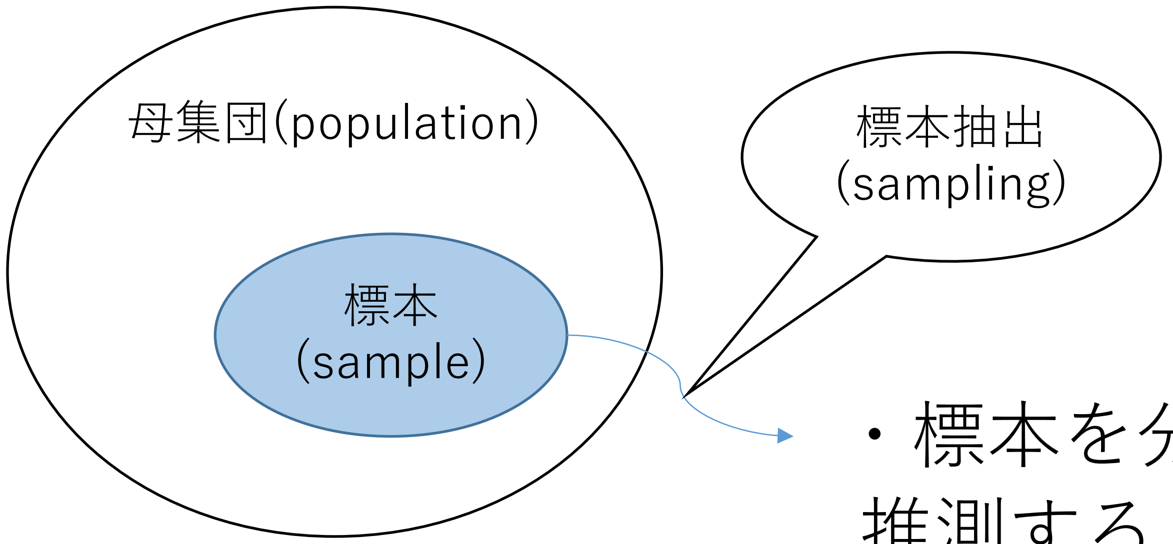
標本分布
と
正規分布からの標本

2020/4/24
B4
佐藤光

目次(Index)

目次(Index)

統計的推測(statistical inference)



・標本を分析することで母集団の特性を推測する。

・標本は*単純ランダムサンプリング*(単純無作為抽出)により選ぶ。

なぜ全体の調査を行わないのか

母集団が大きい場合

全体調査が意味を持たない場合

- デバイス製品の強度検査

現在の測定では測定不可能な要素が必要な場合

- 来年の企業業績予想、経済成長率

統計的推測における確率①

表1貯金額

人物	A	B	C	D	E	
貯金(万)	100	50	25	1000	400	平均315

3人選び貯金額の平均を求める

(ABC)の場合(100+50+25)/3=58.333…

(BCD)の場合(50+25+1000)/3=358.333…

(DEA)の場合(1000+400+100)/3=500となる

→つまり、選び出す標本によって平均値が大きく変化する。

統計的推測における確率②

*統計量*はランダムサンプリングによって抽出される標本によって異なるので統計量は母集団分布に従う確率変数であると考えられる。

また、サンプルの度に統計量が異なることを*サンプリング誤差*と言う。

そして、標本は同一の母集団分布f(x)に従うn個の独立な確率変数

であると考えられる。

統計量…標本から計算されるもの、母集団の様々な推測に用いられる。

母集団の特性を知る時にどのように統計量を使うか?①

母集団を知るために必要な性質

- 母集団の分布は*母平均*μと*母分散*σ^2により規定する
- 母集団の分布が正規分布の時、μとσ^2によって完璧に特性を表すことができる。

母集団の特性を知る時にどのように統計量を使うか?②

→標本平均と標本分散を用いる。

➢標本平均$\bar{X}$は

$$\bar{X}=\frac{X_1+X_2+\cdots X_n}{n}$$

➢標本分散s^2は

$$s^2=\frac{1}{n-1}\{(X_1-\bar{X})^2+(X_2-\bar{X})^2+\cdots(X_n-\bar{X})^2\}$$

と定義される。

(注)標本分散の分母は(n-1)である。

標本平均、標本分散と母平均、母分散との関わり合い①

$E(\bar{X})=E(\frac{X_1+X_2+\cdots X_n}{n})=\frac{n\mu}{n}=\mu$

$V(\bar{X})=V(\frac{X_1+X_2+\cdots X_n}{n})=\frac{n\sigma^2}{n^2}=\frac{\sigma^2}{n}$-(配布資料(a)参照)

→①*標本平均*の期待値がμと一致する

　②分散は1/nに比例する、つまりn→∞の時$\bar{X}$の分散が0に近づいて、$\bar{X}\rightarrow\mu$と定数に収束する。

標本平均、標本分散と母平均、母分散との関わり合い②

$$\mathrm{E}(s^2)=\mathrm{E}\left(\frac{1}{n-1}\{(X_1-\bar{X})^2+(X_2-\bar{X})^2+\cdots(X_n-\bar{X})^2\}\right)$$

$$=\frac{(n-1)\sigma^2}{n-1}=\sigma^2$$ −(配布資料(b)参照)

→*標本分散*の期待値がσ^2に一致する。つまり、母分散に不偏に推定することが可能である。

何故、標本分散の分母をnにしないのか

$$S^2 = \frac{1}{n}\{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots (X_n - \bar{X})^2\}$$

$$\mathrm{E}(S^2) = \frac{n-1}{n}\sigma^2$$

つまり、nの大きさによって標本分散S^2の値が変化する。

例えばn=10の場合、$\mathrm{E}(S^2) = 0.9\sigma^2$となり10％母分散$\sigma^2$を過小評価する事となる。

目次(Index)

標本分布(Sample distribution)

<u>分布の種類</u>

- 二項分布
- ポアソン分布
- 正規分布

ベルヌイ分布

ベルヌイ分布とは…
2つの結果のみが生起する試行(ベルヌイ試行)において
生起するときを1、生起しないときを0の値を取る確率変数Xが
従う分布

$$Pr(X=1)=p, Pr(X=0)=1-p$$

$$E(X)=p$$

$$V(X)=p(1-p)$$

二項分布①

二項分布とは

…生起する確率pのベルヌイ試行をn回行ったときに生起した回数Xが従う分布、一般にB(n,p)と表す。

$$\Pr(X=k) = {}_{n}C_{k}\, p^{k}(1-p)^{n-k}$$

$$E(X) = np$$

$$V(X) = np(1-p)$$

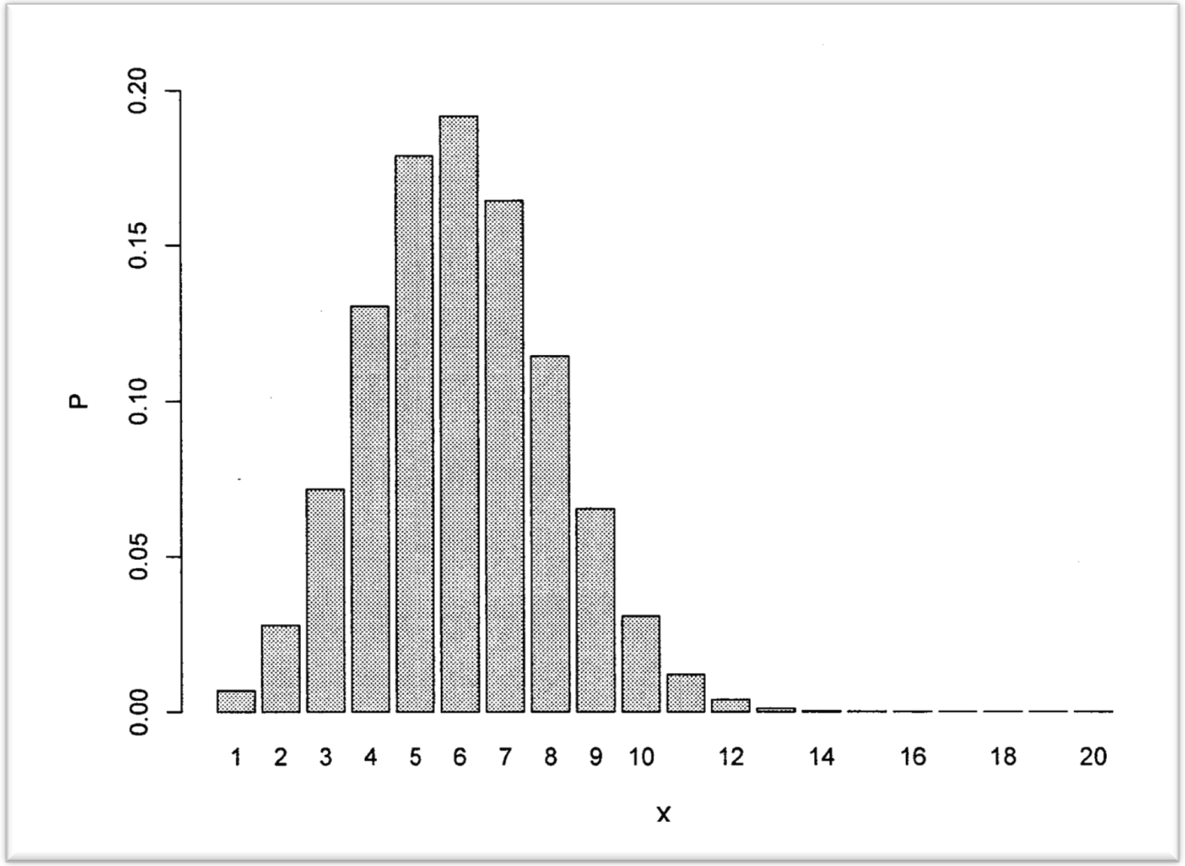
二項分布②

20回試行して、生起確率が0.3の時の確率分布

E(X)=np=6.0

V(X)=np(1-p)=4.2

$$z = \frac{S - np}{\sqrt{np(1-p)}}$$



ポアソン分布①

ポアソン分布とは

…ベルヌイ試行における生起確率pが小さい時に従う分布

二項分布の期待値np＝λとして固定したままn→∞とした時の分布と一致する。

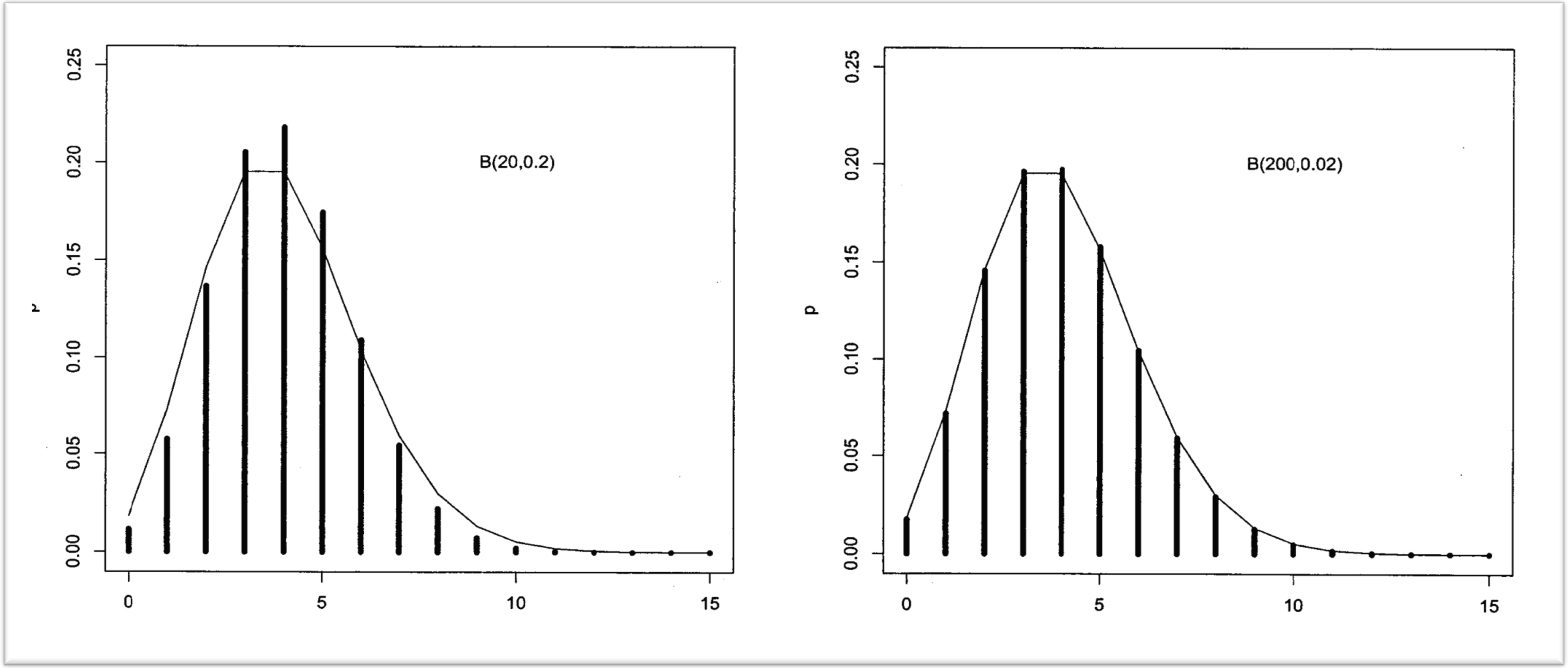
一般にPo(λ)と表す。

$$\Pr(X=k)=\lim_{n\to\infty} {}_{n}C_{k}\, p^{k}(1-p)^{n-k}=\frac{\lambda^{k}}{k!}e^{-\lambda}$$

$$E(X)=\lambda$$

$$V(X)=\lambda$$

二項分布とポアソン分布の比較



目次(Index)

C.F.ガウスによる誤差理論①

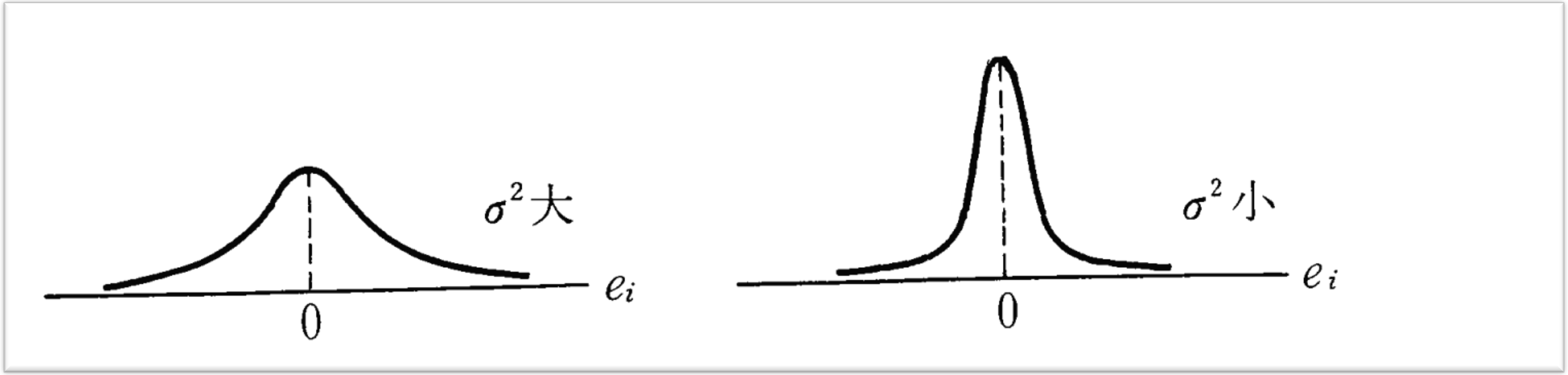
測定による各要素の測定値X_iは

$$X_i = (真の値(\mu)) + (測定誤差(e_i))$$

という2要素を含んでいるはずと考えた。

C.F.ガウスによる誤差理論②

各誤差e_iは平均0,分散σ^2(定数)の正規分布N(0,σ^2)
に従うと仮定した。



なぜ母集団を正規分布と仮定するのか

分析対象となるものが正規分布である場合が多い

変数の変換によって正規分布で表される場合もある

- 体重の1/3乗、収入の対数

*中心極限定理*による正規分布への近似

独立した正規確率変数の和が正規確率変数である

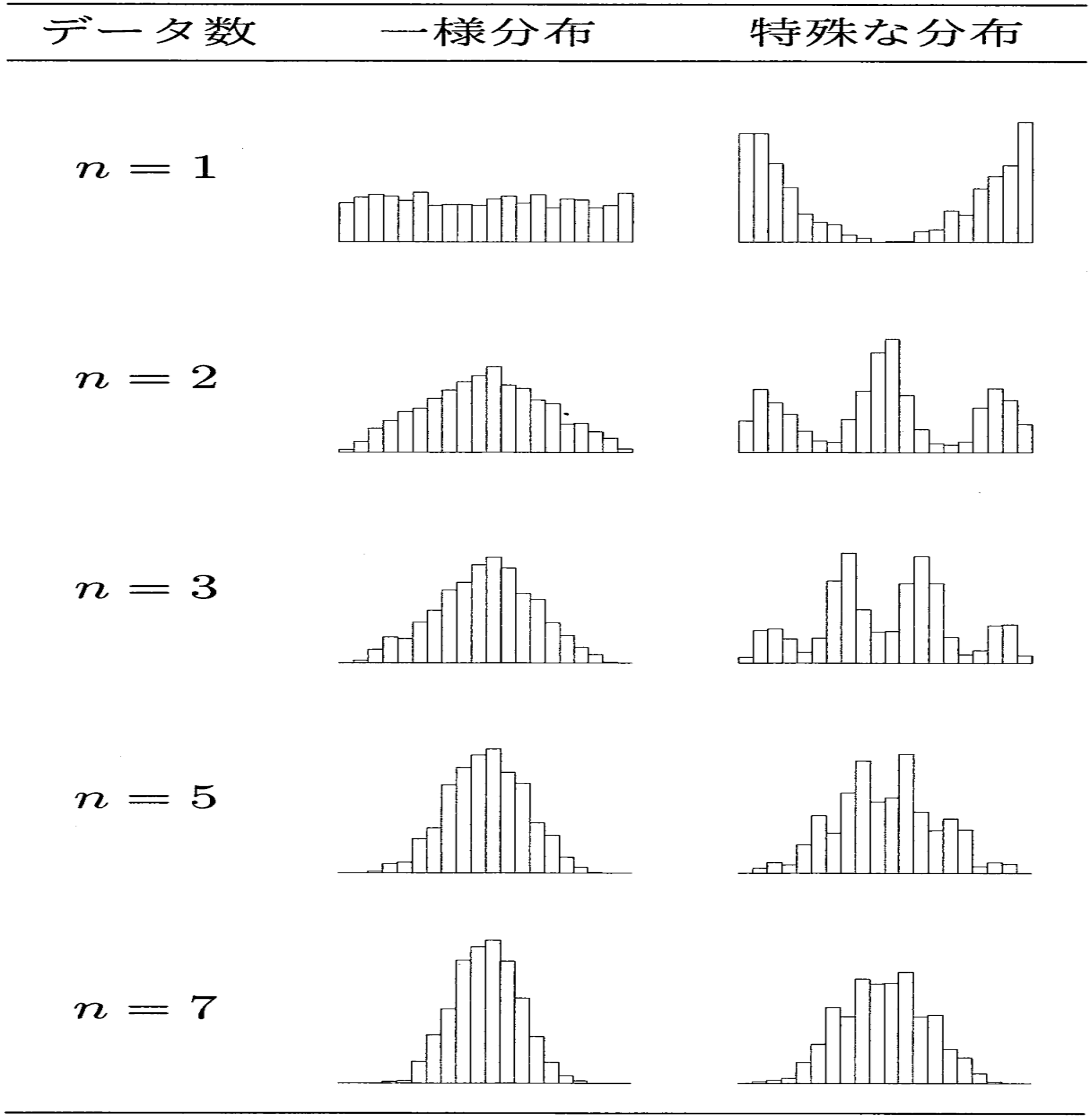
- 正規分布の再生性

中心極限定理①

$$\lim_{n\to\infty} P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

同一の分布に従う確率変数の標本平均の分布が
確率変数の数が大きくなった場合に元の分布に関係なく
正規分布に収束する。

中心極限定理②



実用上はn=5~15で十分

正規分布の再生性

①正規確率変数の線形変換が正規確率変数

- Xが正規分布N(μ ,σ^2)に従う場合、aX+bは N(a μ +b,$a^2\sigma^2$)に従う

②独立した2つ以上の正規確率変数の和や差も同様に正規確率変数となる

- XがN(μ_1,σ_1^2) YがN(μ_2,σ_2^2)従うときX+Yは N($\mu_1+\mu_2,\sigma_1^2+\sigma_2^2$)、X-YはN($\mu_1-\mu_2,\sigma_1^2+\sigma_2^2$)に従い 一般的にaX±bYはN($a\mu_1 \pm b\mu_2, a^2\sigma_1^2+b^2\sigma_2^2$)に従う

分散が既知であるときの$\bar{X}$の標本分布①

N(μ,σ^2/n)に従う$\bar{X}$を標準化した

$$Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

はN(0,1)の*標準正規分布*に従う。

パーセント点(percentage point)…ある点より上側又は下側又は両側となる確率が100α%となる点。一般にZ_αと表す。

→信頼区間や仮説検定に用いる

分散が既知であるときの$\bar{X}$の標本分布②

例題

物の長さの測定でμ =10.0と仮定する。

10回測定で$\bar{X}$=10.05となり、標準偏差0.03とする。

$$Z=\frac{10.05-10}{0.03/\sqrt{10}}=5.270$$

$Z_{0.00005}=3.89$と比較すると仮定は明らかに誤りと分かる。

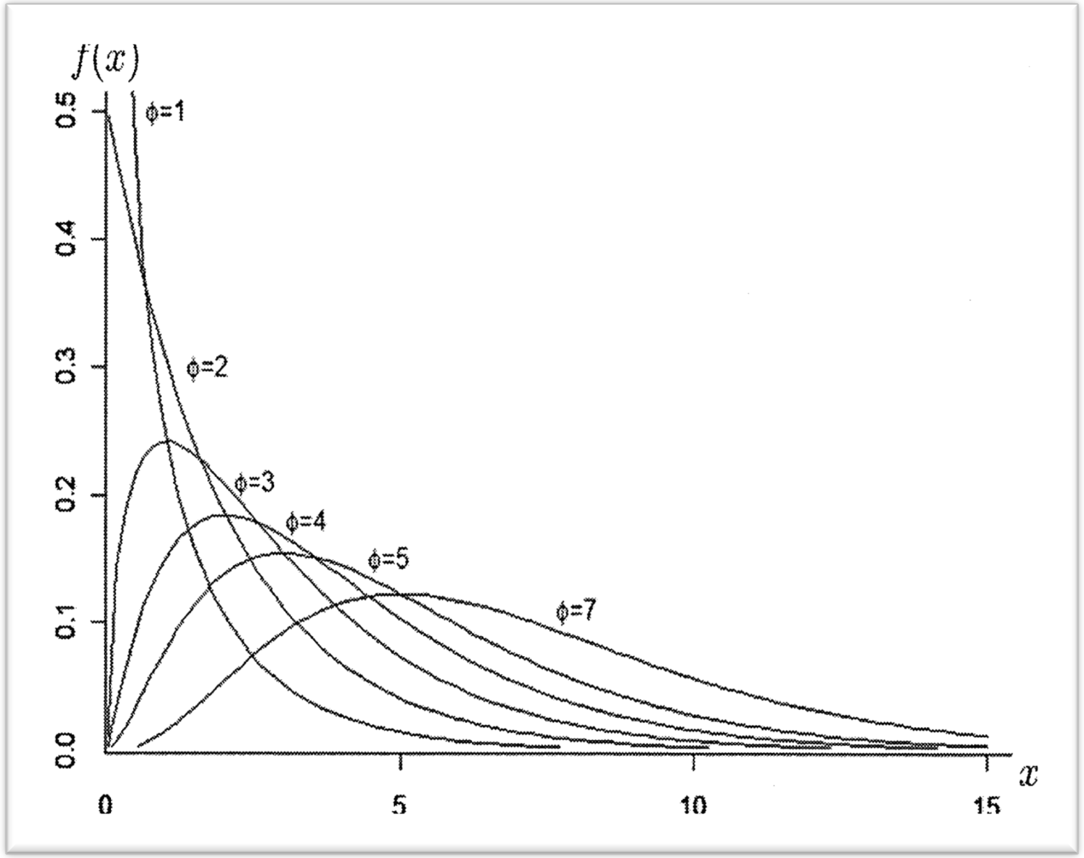
標本分散の標本分布①

自由度kのχ^2分布…$Z_1, Z_2 \cdots Z_k$を独立なN(0,1)に従う確率変数とすると

$$\chi^2 = Z_1^2 + Z_2^2 \cdots Z_k^2$$

と定義する。-(参考資料(c))

$\chi^2 \equiv (n-1)s^2/\sigma^2$と近似されて自由度はn-1である。



標本分散の標本分布②

なぜ自由度がn-1になるのか…

$$(X_1 - \bar{X}) + (X_2 - \bar{X}) + \cdots + (X_{n-1} - \bar{X}) + (X_n - \bar{X}) \equiv 0$$

という制限では、第n変数は第1変数から第(n-1)変数によって完全に1つに決定される。

よって自由に動ける変数は見かけより1つ少なくなる。

標本分散の標本分布③

自由度が1の時のχ^2分布の確率密度関数

$$f(x) = \frac{1}{\sqrt{2\pi x}} e^{-x/2}$$

であり、モーメント母関数は

$$\Phi_1(t) = \int_0^\infty f(x) e^{tx} dx = \frac{1}{\sqrt{1-2t}}$$

である。

標本分散の標本分布④

自由度がnの場合

$$\Phi_n(t) = (\frac{1}{\sqrt{1-2t}})^n$$

→χ^2分布は*再生性*を持つ

$$\Phi_{n1}(t) * \Phi_{n2}(t) = (\frac{1}{\sqrt{1-2t}})^{n1}(\frac{1}{\sqrt{1-2t}})^{n2} = (\frac{1}{\sqrt{1-2t}})^{(n1+n2)}$$

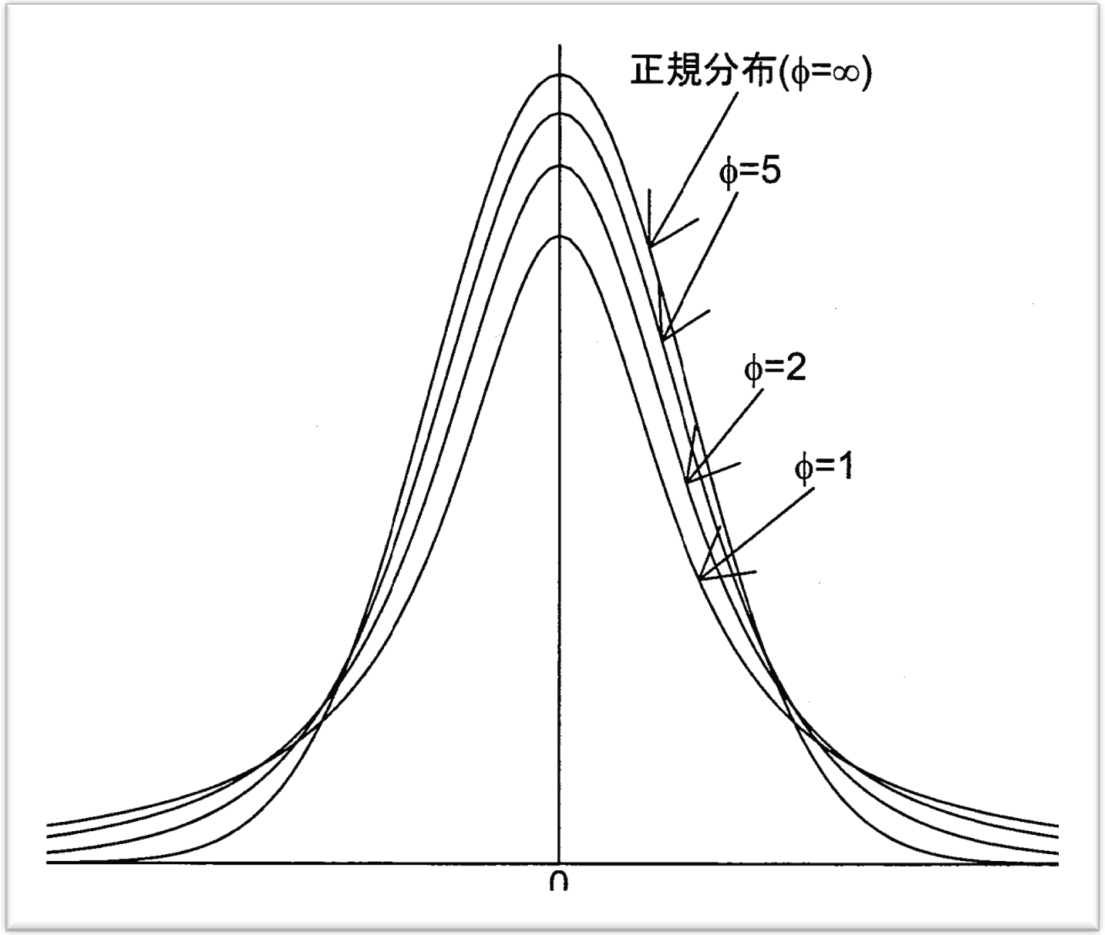
分散が未知の時の$\bar{\mathrm{X}}$の標本分布①

t統計量(Student’s t-statistic)

$$t=\frac{\bar{X}-\mu}{s/\sqrt{n}}$$

を定義する。(分散σ^2が未知なので標本分散s^2と入れ替える)

勿論、N(0,1)には従わない。



分散が未知の時の$\overline{\mathrm{X}}$の標本分布②

t分布を式変形すると…

$$t=\frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\frac{s}{\sigma}}=\frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)s^2}{\sigma^2}\Big/_{(n-1)}}}$$

→分子は標準正規分布N(0,1)に従い分母は自由度(n-1)の
χ^2分布に従う。更に$\overline{\mathrm{X}}$とs^2は独立なので2つの分布の組み合わせ
によってt分布が決定される。

2標本問題(two-sample problem)

2つの母集団の比較を扱う問題、その母平均の差$\mu_1 - \mu_2$

を分析することが多い。

$\Rightarrow \bar{X} - \bar{Y}$の大小をそれぞれの分散を考慮して分析する。

- それぞれの母分散$\sigma_1^2 \sigma_2^2$が既知のとき
- 母分散が未知だが等しいとき($\sigma_1^2 = \sigma_2^2 = \sigma^2$)
- 母分散が未知のとき

母分散が既知のとき

→$\bar{X}-\bar{Y}$はN$(\mu_1-\mu_2, \sigma_1^2/m+\sigma_2^2/n)$に従う。
標準化した

$$Z=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\left(\sigma_1^2/_m\right)+\left(\sigma_2^2/_n\right)}}$$

は、N(0,1)に従い、このZを用いることで推定や検定を行う。

母分散が未知だが等しいとき $(\sigma_1^2 = \sigma_2^2 = \sigma^2)$

→*2標本のt分布*(two-sample t-statistic)に従う

$$s^2 = \frac{\sum_{i=1}^{m}(X_i-\bar{X})^2+\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{m+n-2}$$

標準化すると

$$t=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{s\sqrt{\left(\frac{1}{m}\right)+\left(\frac{1}{n}\right)}}$$

自由度が(m+n-2)のt分布に従う。⇒2標本のt検定

母分散が未知のとき

→ウェルチの近似を用いてt分布に帰着する

ウェルチの近似法…

$$t=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{m}+\frac{s_2^2}{n}}}$$

は近似的に自由度が

$$\gamma=\frac{\left(\frac{s_1^2}{m}-\frac{s_2^2}{n}\right)^2}{\frac{\left(s_1^2/m\right)^2}{m-1}+\frac{\left(s_2^2/n\right)^2}{n-1}}$$

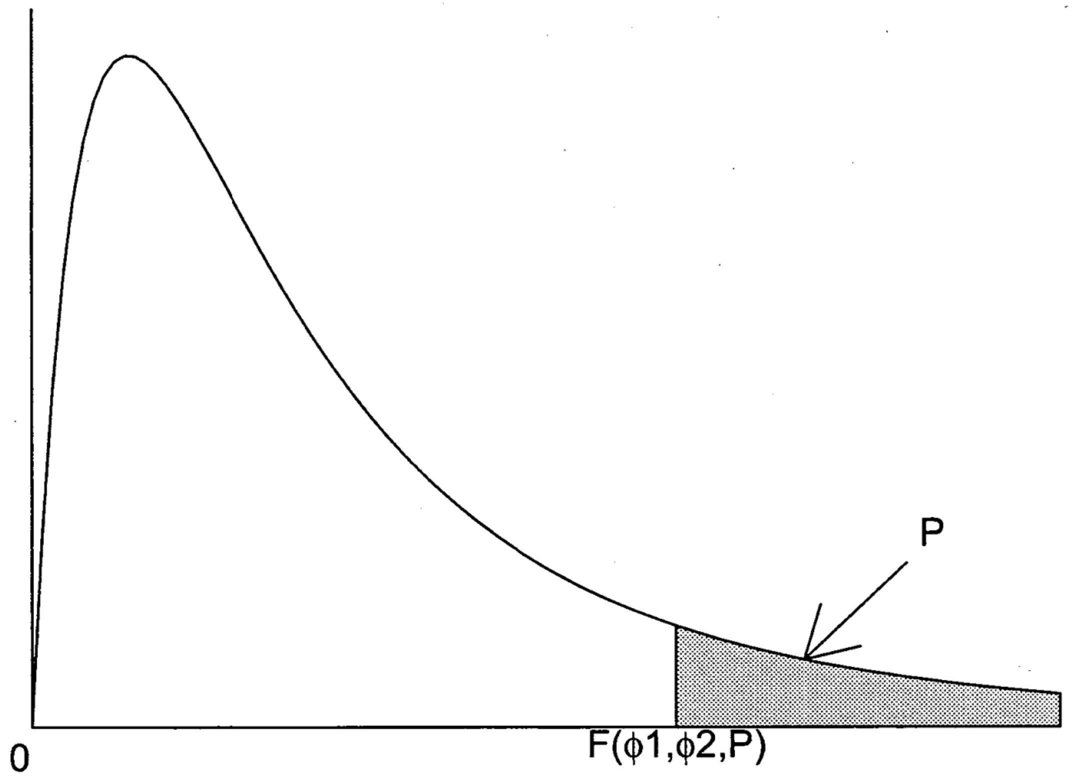
に最も近い整数γ^*のt分布$t(\gamma^*)$に従う。

標本分散の比の標本分布①

⇒s_1^2/s_2^2 ~1ならば母集団の$\sigma_1^2/\sigma_2^2=1$
と推測できる為
→標本分散の比は*F分布*に従う。
F分布…χ^2分布に従う確率変数の比

$$F=\frac{U/k_1}{V/k_2}$$

と定義される。



標本分散の比の標本分布②

s_1^2とs_2^2に関して…

$$F=\frac{\dfrac{(m-1)s_1^2}{\sigma_1^2}\Big/(m-1)}{\dfrac{(n-1)s_2^2}{\sigma_2^2}\Big/(n-1)}=\frac{\sigma_2^2}{\sigma_1^2}\cdot\frac{s_1^2}{s_2^2}$$

自由度(m-1,n-1)に従うF分布となる。

一般にF=${}^{s_1^2}/_{s_2^2}$の形で用いられる。

標本相関係数の標本分布①

→フィッシャーのz変換(Fisher's z-transformartion)を用いる。

母共分散$\sigma_{XY}=E(X-\mu_X)(Y-\mu_Y)$

母相関係数$\rho_{XY}=\dfrac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$

標本共分散$s_{XY}=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$

標本相関係数$r_{XY}=\dfrac{s_{XY}}{s_X \cdot s_Y}$

標本相関係数の標本分布②

⇒母相関係数ρ_{XY}に対しての標本相関係数r_{XY}の標本分布を求める。

$$z=\frac{1}{2}\log\frac{1+r_{XY}}{1-r_{XY}}$$

$$\eta\,(イータ)=\frac{1}{2}\log\frac{1+\rho_{XY}}{1-\rho_{XY}}$$

と変換する。zの標本分布はnが大きい時にN(η,1/(n-3))に従う。

標準化すると…

$$\sqrt{n-3}(z-\eta)$$

はN(0,1)に従う。

有限母集団の標本分布

- Nが有限であることを含む修正

→①期待値E($\bar{X}$)は母平均μと一致する

　②分散V($\bar{X}$)は

$$V(\bar{X})=\frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}$$

となる。

⇒分散は有限母集団修正を考慮する必要がある。

本日のまとめ

統計的推測のアウトライン	• →母平均、母分散と標本平均、標本分散による統計的推測
分布の種類	• →二項分布、ポアソン分布、正規分布の性質
正規分布による分析	• →χ^2分布、t分布そしてF分布への派生

本日のまとめ(1標本問題)

母分散が既知のときの$\bar{X}$の標本分布	母分散が未知のときの$\bar{X}$の標本分布	標本分散s^2の標本分布
• 正規分布を考える。	• t分布→(正規分布/χ^2分布)	• χ^2分布

本日のまとめ(2標本問題)

母分散が既知のとき の $\bar{X}-\bar{Y}$の標本分布	母分散が未知だが等しいとき$\bar{X}-\bar{Y}$の標本分布	母分散が未知のときの $\bar{X}-\bar{Y}$の標本分布
• 正規分布を考える。	• 2標本のt検定	• ウェルチの近似法→近似的にt分布

本日のまとめ(2標本問題)

標本分散の比の標本分布

- F分布

相関係数の標本分布

- フィッシャーのz変換