

第 3 回 正規分布

★ 教材「生物統計学__母集団と標本 2013」を予習しながら空所を埋めておくこと

A. 母集団と標本

1. 部分（標本）から全体（母集団）を推測

第 1 回の授業では池にすむ魚の平均体重を調べるために、池の中の魚の一部を捕獲して、そこから推測することの是非について論じた。このように部分（統計学では標本という）から全体（母集団という）を推測することはよく行われることである。しかし、部分から全体への一般化はどれほど危険を伴うことだろうか？以下の例を考えてみよ。

予習課題：クロールのタイムがほぼ同等であった日本人の 15 歳の男子 40 人を均等に 2 つのグループに分け、一方の 20 人に新しいクロールの教え方を適用したところ、従来の教え方を適用したグループ 20 人よりも 50m のタイムが向上した。

では、この結果は

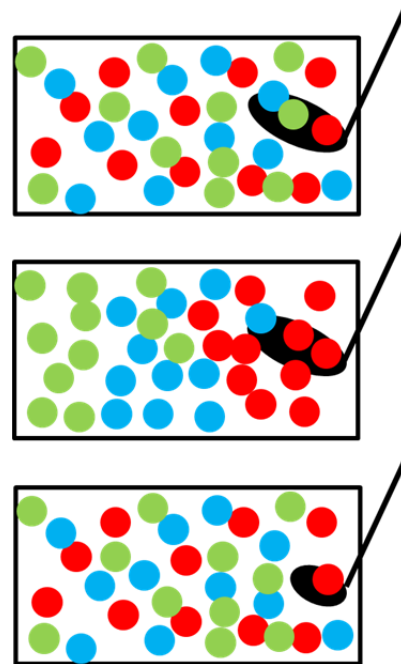
- ① 日本人の 15 歳の男子全体, ② 日本人の 13~16 歳の男子全体, ③ 日本人の 15 歳男女とも, ④ 日本人男子全体, ⑤ アメリカ人の 15 歳の男子全体, ⑥ 砂漠に住む 15 歳の男子全体,
- ⑦ すべての人間・・・に適用できるだろうか？

このようにある事実あるいはある実験結果をどこまで推論を拡張してもよいだろうか？

予習で自分で考えて、以下に書いておいてください。

★ 自分の考え

★ 授業でのグループでの議論



2013 年 10 月 22 日

予習課題：駆け出しの研究者である A 君は宍道湖の北東の岸辺でシジミを 20 匹，一度に網で捕まえて，それを 10 匹ずつに分けた．一方に薬 A を与え，一方には対照として何も与えなかった．その結果，薬 A を与えた方が成長がよかった．

では，この結果は

① 宍道湖すべてのヤマトシジミ，② 日本のヤマトシジミ，③ 日本のシジミ類，④ 日本の貝類，⑤ アジアのシジミ類，⑥ 世界の貝類，⑦ すべての生物⑧ 火星で陸上養殖したヤマトシジミに適用できるだろうか？ いろいろ想像してみよう．

このようにある事実あるいはある実験結果をどこまで推論を拡張してもよいだろうか？

予習で自分で考えて，以下に書いておいてください．

★ 自分の考え

★ 授業でのグループでの議論

2. 記述統計と推測統計

() 統計 観測値を要約し，記述するための方法

() 統計 観測値を推定や予測をするための基礎として用いる方法．

すなわちこれまでに観測されていない状況について推測する方法

農学研究ではおもに推測統計に興味が置かれる．では記述統計と推測統計の違いは対象をどうとらえることによって生じるのか？

①	日本の人口は 2004 年現在、1 億 2 千万人強である。	記述統計	・	推測統計
②	日本の人口は 2050 年には 1 億人になる。	記述統計	・	推測統計
③	大山に 2007 年 2 月 1 日に 29cm の降雪があった。	記述統計	・	推測統計
④	大山には冬、雪が降る。	記述統計	・	推測統計
⑤	この風邪薬はよく効く。	記述統計	・	推測統計
⑥	この肥料をブロッコリーに施肥したところ、ビタミン C が 50% 増加した。	記述統計	・	推測統計
⑦	この肥料はブロッコリーの品質を向上させる。	記述統計	・	推測統計

3. 母集団と標本とは？

標本

4. 実在母集団と仮説的無限母集団

- ② () 母集団
母集団に属する要素が無数にある母集団
例：工場で作られる製品、医薬品の効果を調べる対象であるヒト

よくある誤解に母集団に対して何パーセント調べるとよいという人がいる。これは間違いである。標本数だけで（母集団が無量大であろうと），ほとんど精度が決定することが統計学で証明されている。

次の①から⑥について、それぞれ母集団は何か？標本は何か？母集団は実在母集団か、仮説的無限母集団かを考えよ。予習で考えておくように。

- ① ある市の市長選で有権者 500 名を対象にだれに投票するかを調査した。
母集団 () 標本 ()
実在母集団・仮説的無限母集団
- ② あるテレビ局の番組の視聴率について、20 歳代 300 人に尋ねた。
母集団 () 標本 ()
実在母集団・仮説的無限母集団
- ③ ある食品工場では、工場排水の水質保証のために 1 ヶ月に 6 回の割合で放流口から採水して排水分析をしている。
母集団 () 標本 ()
実在母集団・仮説的無限母集団
- ④ 羊に与えるえさに混ぜると成長がよくなるという薬を 5 匹の羊に与え、調査した。
母集団 () 標本 ()
実在母集団・仮説的無限母集団
- ⑤ ある製薬会社は血圧を低下させる薬の候補である A を 20 匹のラットに与え、効果があることを明らかにしたので人間への応用を考えている。
母集団 () 標本 ()
実在母集団・仮説的無限母集団
- ⑥ ある家電メーカーでは、エアコン製造ラインで種々の不良が発生している。そこで工場から 200 枚の基盤を調べたところ、11 枚の不良品を発見した。
母集団 () 標本 ()
実在母集団・仮説的無限母集団

★ 統計解析とは標本を調べることによって母集団を推定するのが目的である。
しかし、母集団と標本の関係は理論では明確であっても現実にはそう簡単ではない。

サンプリングのパラドックス（新・涙なしの統計が P17 参照）ということばがある。標本はそれが母集団を代表していない限り、誤った理解を導くものとなる。しかし、その標本をとる前に、母集団がわかっているわけではなから、標本を正しく（無作為に）とったといえるだろうか？中身の分からない母集団から、適切に標本をとれるはずがあろうか？

さらにある医者が高血圧の患者が何名かやってきた。そこから高血圧の患者全体への推論ができるだろうか？医者に来る患者に何らかの偏りがあるかもしれない。

質問 それぞれの医者 of 事例についてどんな偏りが考えられるだろうか？

- ① 高血圧専門の医者なら・・・
② 小児科医なら・・・

この医師は自分の標本（やってきた患者）にふさわしい母集団を考えることになる。

★ 自分の考え あなたがもし医者だったら（条件をつけてもよい）何を母集団と考えるだろうか？（標本は自分のところにきた患者）

★ 授業でのグループでの議論

★ 標本と母集団の不一致

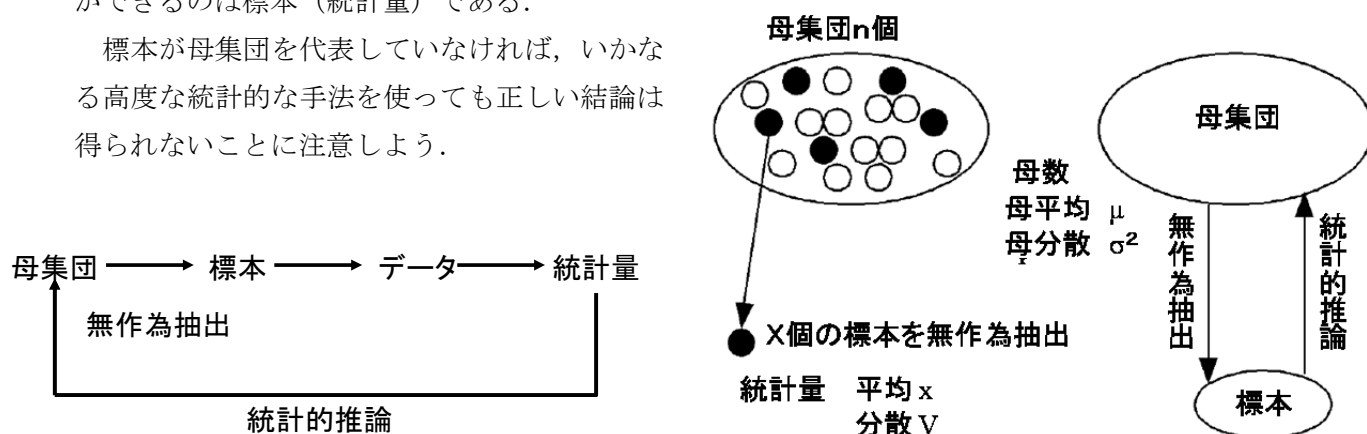
選択バイアス

アメリカ空軍では第 2 次世界大戦のとき，戦闘機が攻撃から帰ってきたら，攻撃を受けた場所を調査した．その結果，A：ほとんど間違いなく被弾したところ，B：ほとんど被弾していなかったところがあった．A と B のどちらを補強したらよいだろうか？

（ヤバい経営学 ヴァーミューレン 東洋経済新報社から）

母集団の特徴を表す数値を母数（parameter）という．母平均 μ ，母分散 σ^2 など，ギリシャ文字を使用する．これに対して標本のそれを統計量（statistic）という．標本平均 m ，標本分散 V などローマ字を使用する．統計解析とは母数を推定するために集めた標本の統計量から母数を推定することである．わたしたちが知りたいのは母集団（母数）である．しかし，実際に知ることができるのは標本（統計量）である．

標本が母集団を代表していなければ，いかなる高度な統計的な手法を使っても正しい結論は得られないことに注意しよう．



4. 標本の抽出

★ 母集団を代表するように標本を抽出しなければならない.

★ 統計学は標本を無作為に抽出することを要請する.

無作為 (at random) にサンプリングしなければならない 注意! でたらめとはちがう

無作為とは母集団のどの要素も等しい確率で抽出されることをいう

乱数表, さいころ, 乱数さい (正二十面体のさいころ), トランプ, コンピュータの疑似乱数

無作為のあり方も専門的知識が必要である.

★ 無作為抽出の例

- ① 松江市市長選を予想するために乱数表を使って, 500 人の有権者を抽出した.
- ② K牧場の牛 250 頭のうち, 乱数さいで 10 頭を選んで, 血液検査した.
- ③ 瓶詰めの牛乳の脂肪価の検査で, よくかき混ぜた後ピペットを縦に入れ, 上層から下層までをサンプリングした.
- ④ ベルトコンベアで運ばれる製品をランダムな時間間隔でサンプリングした.

★ 教材「生物統計学__正規分布の特徴 2013」を予習しながら空所を埋めておくこと

B. 正規分布

1. 正規分布の特徴

- ① 母平均 μ と母分散 σ^2 を与えると形が決まる. これを $N(\mu, \sigma^2)$ と書いて表現する.
- ② 平均 μ を中心にして左右対称である. よって, 平均より大きい値あるいは小さい値を取る確率はそれぞれ (,) である.
- ③ 曲線は平均 μ の近傍で高く, 両側に行くにしたがって単調に低くなる.
- ④ 平均 μ は曲線の位置を決める. 平均 μ のみ異なる 2 つの曲線は左右に移動させれば重ねることができる (図 2).
- ⑤ 標準偏差 σ は曲線の形を決める. σ が大きければ曲線は扁平になる (図 3).
- ⑥ (a) $\mu - \sigma$ と $\mu + \sigma$ の間の確率変数を取る確率は約 0.683 である (図 4).
(b) $\mu - 2\sigma$ と $\mu + 2\sigma$ の間の確率変数を取る確率は約 0.954 である.
(c) $\mu - 3\sigma$ と $\mu + 3\sigma$ の間の確率変数を取る確率は約 0.997 である.
- ⑦ 0.95 (95%) の確率で $\mu - 1.96\sigma$ と $\mu + 1.96\sigma$ の間の確率変数を取る
0.99 (99%) の確率で $\mu - 2.576\sigma$ と $\mu + 2.576\sigma$ の間の確率変数を取る

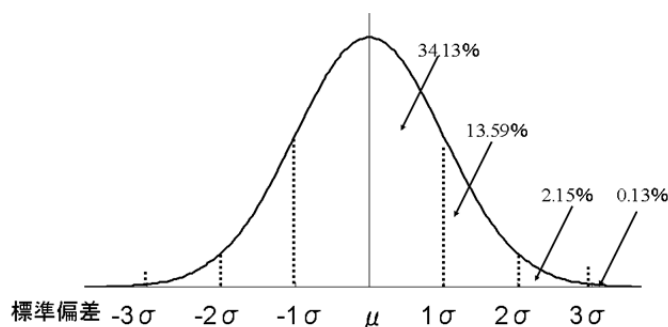


図 1 正規分布

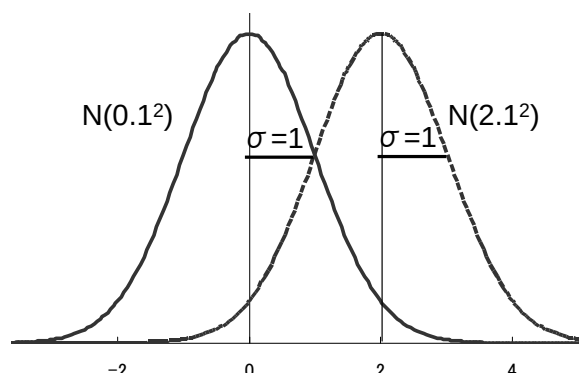


図 2 正規分布 $N(\mu, 1^2)$ の確率密度関数

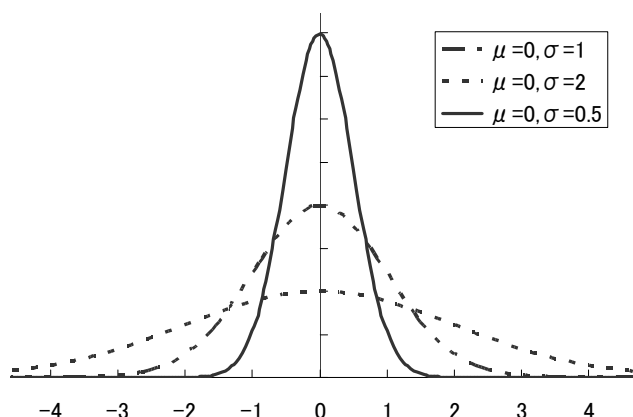


図 3 正規分布 $N(0, \sigma^2)$ の確率密度関数

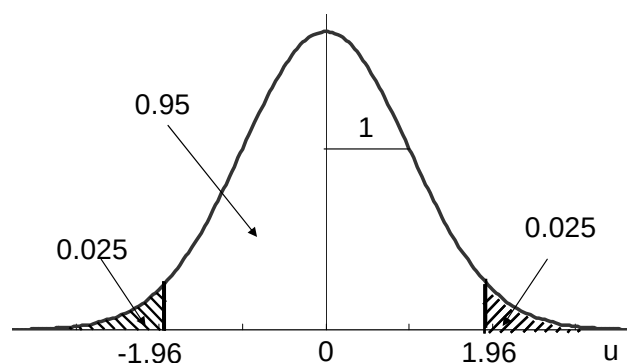
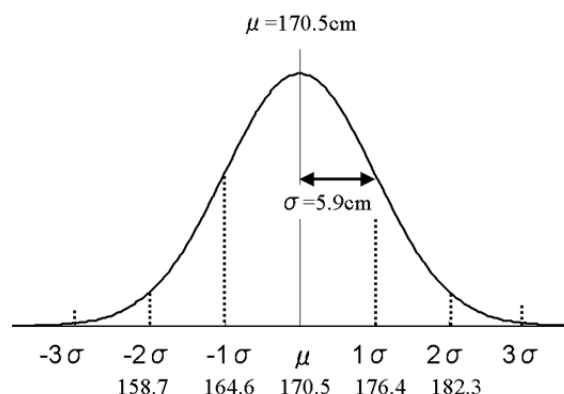


図 4 u の分布, $N(0, 1^2)$

正規分布の例：20～24 歳の男性の身長は人間生活工学研究センターの調査（1992-1994）によると平均 170.5cm，標準偏差 5.9cm であった．身長の分布が正規分布するなら， 2σ 以上平均より背の高い人，すなわち 182.3cm 以上は全体の 2.28%である．平均から標準偏差以内，すなわち 164.6～176.4cm に全体の約 68%が属する．全体の 95%は 158.9～182.1cm に属する．



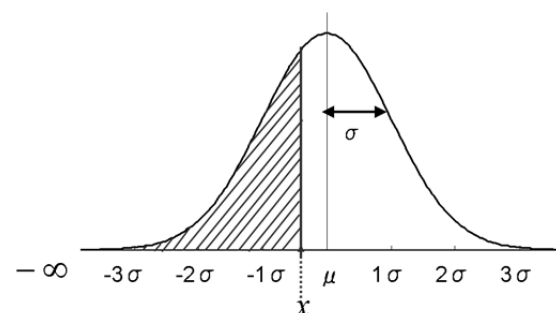
予習での練習：30 歳代の男性の身長の平均は 169.5cm，標準偏差は 5.8cm であった．身長の分布が正規分布するなら，平均から標準偏差以内，すなわち（ ）～（ ）cm に全体の約（ ）%が属する． 2σ 以下平均より背の低い人，すなわち（ ）cm 以下は全体の（ ）%である．全体の 95%は（ ）～（ ）cm に属する．（エクセルファイルで答えを提出してください）

2. 正規分布において任意の値と任意の値の間の範囲をとる確率をエクセルから計算する方法

連続分布であるから，正規分布において任意の値を取る確率は 0 である．任意の値と任意の値の間の範囲を取る確率を計算するにはエクセルの関数を利用するのが簡単である．

エクセルの正規分布に関する関数はいくつかある．今回，利用する NORMDIST 関数は平均 μ ，標準偏差 σ の正規分布において， $-\infty$ （無限大）から x までの値を取る確率を以下のように入力することで計算する．

$$= \text{NORMDIST}(x, \mu, \sigma, \text{true})$$



例えば、前述の身長に関する正規分布の例（平均 170.5cm，標準偏差 5.9cm）で、164.6cm 以下の身長の割合は以下の式で求められる。

=NORMDIST(164.6,170.5,5.9,TRUE)	0.158655
---------------------------------	----------

あるいは平均，標準偏差さえ入力すれば計算するように作られたセルを利用してもよい（なおできるだけエクセルの関数に習熟するためには自分で数式を入れてみる方がよい）。

1. ある値以下となる確率を計算したいとき			
平均		平均	170.5
標準偏差		標準偏差	5.9
ある値		ある値	164.6
確率		確率	0.158655
確率	#NUM!		

予習での練習 20 歳代の男性の身長の平均は 170.5cm，標準偏差は 5.9cm であった。身長が正規分布するならば，160 cm 以下には全体の約（ ）%が属する。175 cm 以下には全体の約（ ）%が属する。（エクセルファイルで答えを提出してください）

エクセルの関数

160cm 以下

175cm 以下

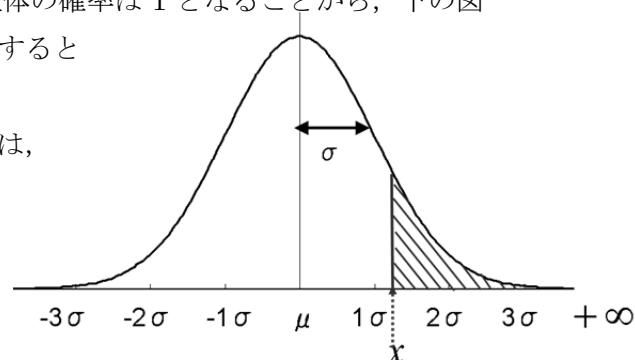
ある値より大きくなる確率を計算するには，正規分布全体の確率は 1 となることから，下の図のように考えて，1 から下の図の斜線部分の確率を引き算すると

$= \text{NORMDIST}(x, \mu, \sigma, \text{true})$ であるから，

すなわち，斜線部に属する確率（ x より大きくなる確率）は，

$= 1 - \text{NORMDIST}(x, \mu, \sigma, \text{true})$

として，計算する。



あるいは平均，標準偏差さえ入力すれば計算するように作られたセルを利用してもよい。

2. ある値以上となる確率を計算したいとき			
平均			
標準偏差			
ある値			
確率	#NUM!		

予習での練習 20 歳代の男性の身長は平均は 170.5cm, 標準偏差は 5.9cm であった. 身長が正規分布するならば, 173 cm 以上には全体の約 () %が属する. 162 cm 以上には全体の約 () %が属する. (エクセルファイルで答えを提出してください)

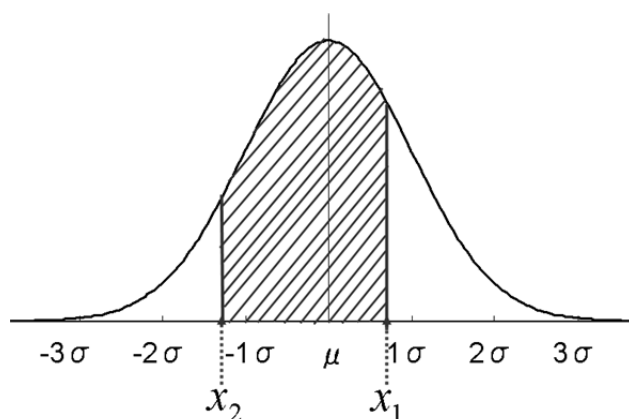
エクセルの関数

173cm 以上

162cm 以上

ある値 (x_2) からある値 (x_1) をとる確率を計算するには, $-\infty$ から x_1 までを取る確率から $-\infty$ から x_2 までを取る確率の差を取る. すなわち下の図のように計算する. エクセルでは

$$= \text{NORMDIST}(x_1, \mu, \sigma, \text{true}) - \text{NORMDIST}(x_2, \mu, \sigma, \text{true})$$
 として, 計算する.



3. AからBの間となる確率を計算したいとき

平均			
標準偏差			
A			
B			
確率	#NUM!		

予習での練習 20 歳代の男性の身長は平均は 170.5cm, 標準偏差は 5.9cm であった. 身長が正規分布するならば, 身長が 160~175cm の間にある人は全体の約 () %である. (エクセルファイルで答えを提出してください)

エクセルの関数

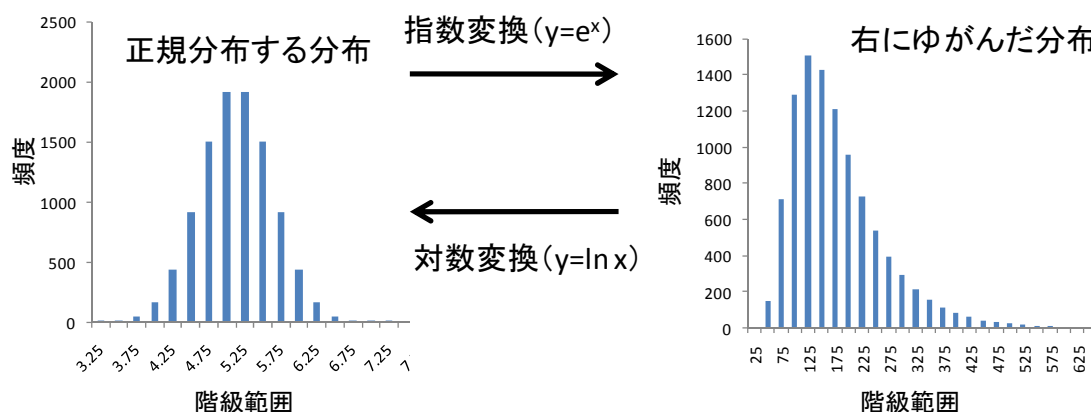
授業での演習問題

A 農園では副業として、カブトムシを飼育している。ある会社と契約を結んで、20g 以上のカブトムシを納入している。20g 以上であれば 1 匹あたりの値段は変わらない。現在の飼育方法だと、カブトムシの平均体重は 25g、その標準偏差は 5g である。

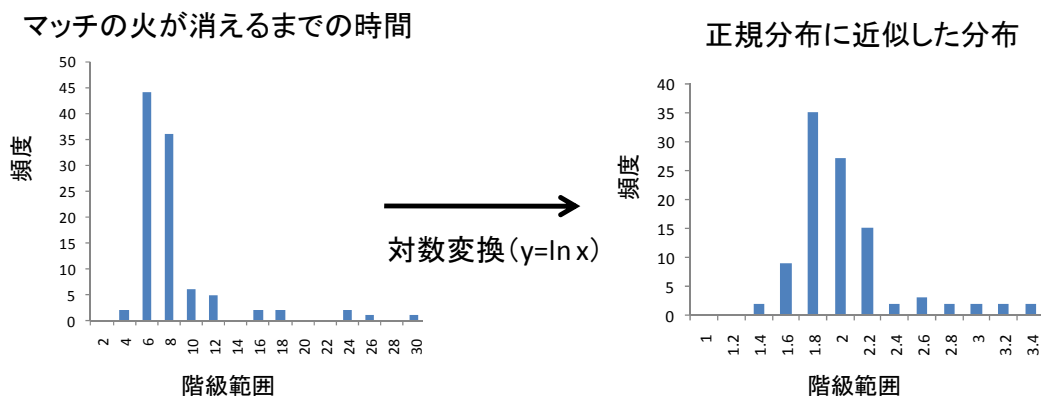
B 君は A 農園に新しい飼育方法を導入すれば、さらにもうかると売り込んできた。その方法を飼育すると、平均は 27g に増えるという。標準偏差を聞くと、B 君は「平均さえ増えればいいじゃない」と答えながらも、標準偏差は 8g であるとのこと。さてこの飼育方法を採用すべきだろうか？

(おまけ 時間があれば授業します) C. 正規分布への変換

母集団が正規分布しない場合でも、数値を変換することによって正規分布に近似できることがよくある。正規分布する分布について、その個々の数値を指数変換 ($y=e^x$) すると右にゆがんだ分布を得ることができ、このような右にゆがんだ分布は年収、貯蓄などの分布にみられる。したがって、年収のように右にゆがんだ分布では、対数変換してやることによって、正規分布に近似できることがある。



例：マッチの火が消えるまでの時間を測定した結果のヒストグラム（左）であり、ここから対数変換すると正規分布に近似できそうであると考えた。各測定値を対数変換した結果、右のヒストグラムを得た。ほぼ正規分布しているとみなし、この対数変換後の値について平均、標準偏差を計算したところ、平均=1.92、標準偏差=0.39 となった。



以上の結果を利用すると、マッチの火が 10 秒以内に消える割合は、10 を対数変換し、2.30 を得る。ここからその割合は 84% であると計算できる。

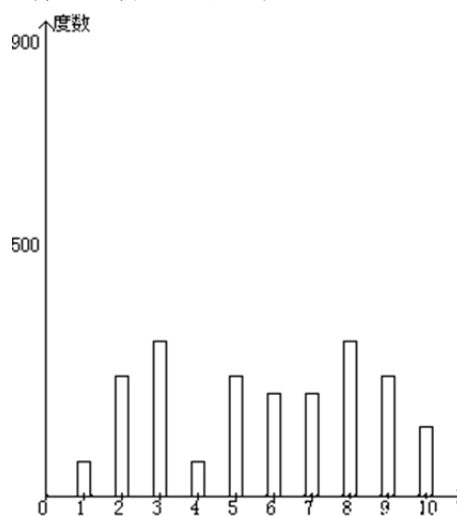
★ 第 6 回以降に学ぶ t 検定、分散分析など多くの基本的な統計手法は母集団が正規分布することが前提である。したがって、数値変換して、正規分布に近似できるかどうかは重要なことである。数値変換の仕方としては、対数をとる、平方根をとる、逆数をとるなどの方法がある。

平均、標準偏差、指定された範囲を入力するとその範囲をとる確率を計算するように以下のシートは作ってあります。適当なセルへコピーして使用してください。				
1. ある値以下となる確率を計算したいとき				
平均	1.92			
標準偏差	0.39			
ある値	2.302585			
確率	0.836701			

D. 標本平均の分布（中心極限定理） ★これはむずかしいですが、とても重要なところ

1. 母集団から無作為に取り出した標本の性質（サンプリング変動）

標本を 5 つ無作為に取るという試行を 2000 回繰り返すと 2000 個のデータからなる標本平均はどんな分布となるだろうか？下のグラフに標本平均の分布を想像して書いてみよう。



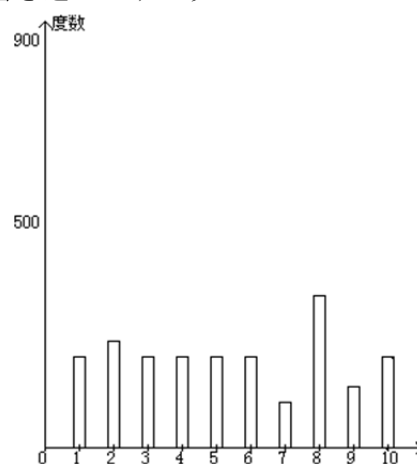
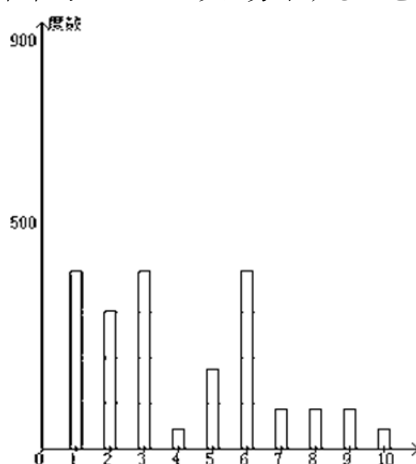
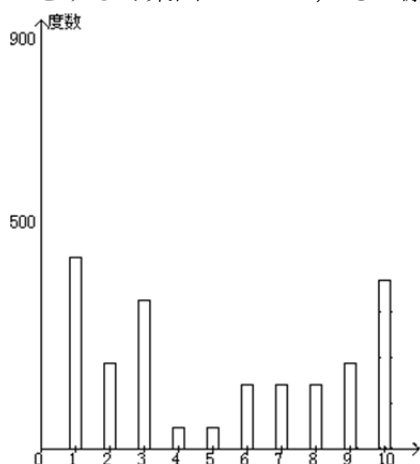
このような標本平均の分布の特徴を考えてみよう。

① 母集団の分布と標本の分布の形はどのように対応するだろうか？

② 標本の平均は母集団の平均に対してどのように位置づけられるだろうか？

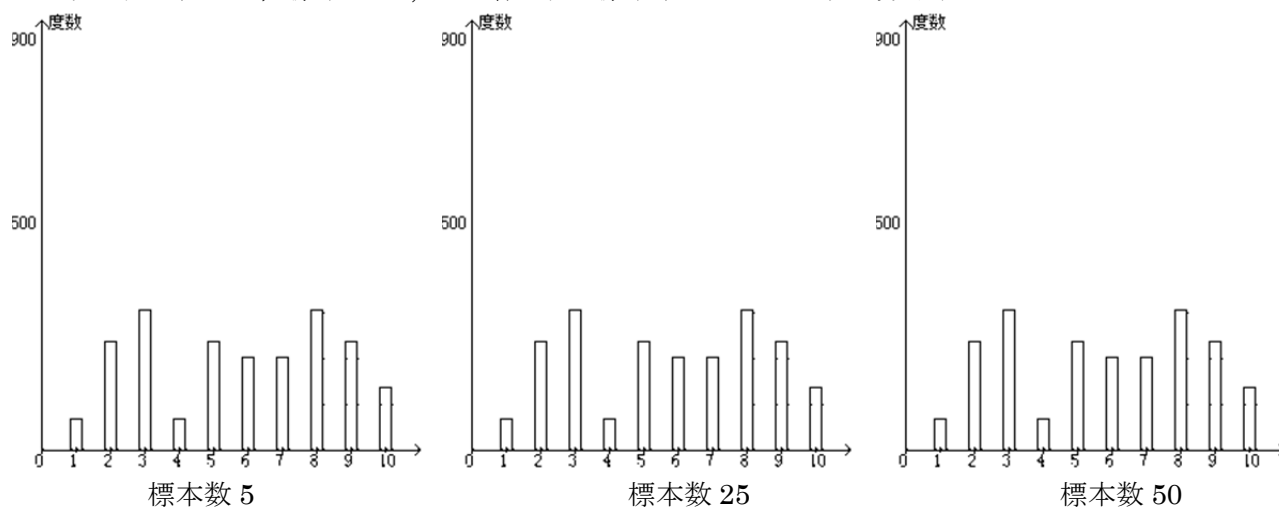
③ 標本の平均はどのくらいばらつくだろうか？

母集団の形と標本平均の分布の形にはどのような関係があるだろうか？下の 3 つの異なる分布をする母集団について、その標本平均がどのように分布するかを書き込んでみよう。



★ 母集団の分布の形にかかわらず、標本平均の分布は（ ）に近似できる。

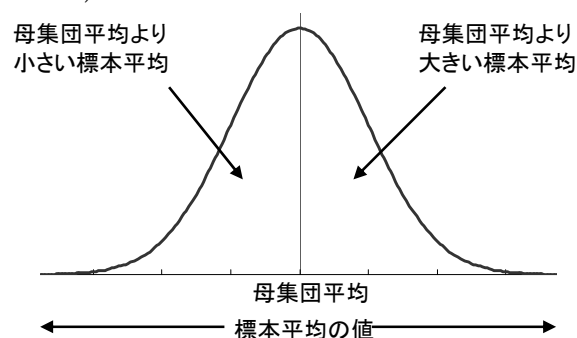
次に標本を取る数を増やしてやるとどうなるだろうか？ すなわち標本を 5 つ無作為にとるという試行に代えて、標本を 25, 50 と増やすと標本平均はどのように分布するか？



★ 標本数を増やすと、標本平均の分布の分散が（大きく・小さく）なる。

以上のことから、標本数が十分に多ければ、母集団の分布によらず、標本平均の分布について、次のことがいえる。

- ① 母集団の分布にかかわらず、標本平均の分布は（ ）に近い形となった。
- ② 標本の平均は母集団の平均とほぼ（ ）。
- ③ 標本平均の分布は、標本数が（多い・少ない）と大きくばらつき、標本数が（多い・少ない）とあまりばらつかなかった。



わたしたちは統計的な調査をするときに、個々のデータそのものに関心を持つことはあまりない。さらに（母集団 ・ 標本）ではなく、（母集団 ・ 標本）に関心を持つ。そして（母集団 ・ 標本）の（平均 ・ ばらつき ・ ヒストグラム）に関心を持つことが一般的である。

前述したように母集団が正規分布に近似しない場合でも、標本数を十分にとれば、その標本平均に関する分布は正規分布に近似できるという特徴がある。わたしたちは母集団の平均に一般的に関心を持つために、たとえ母集団そのものが正規分布していなくても、正規分布には有用性が大きいことがわかる。

E. 予習と宿題

予習と宿題は <https://moodle.cerd.shimane-u.ac.jp/moodle/> をご覧ください。