2019 課題研究基礎 統計学②（正規分布を使って誤差を見る） 事前プリント

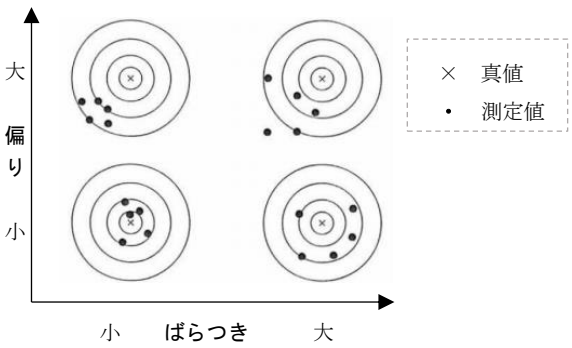
1　測定における誤差には、2 種類ある

測定における誤差とは、測定値－真値 のことです。誤差は「偶然による誤差」と「系統的な誤差」に大別できます。

「偶然による誤差」	測定者のランダムな測定むらなど、偶発的原因で起こるもの	統計的処理で発見可能
「系統的な誤差」	使用した器具の個性、測定者の一定の癖などにより起こるもの	気づきにくい

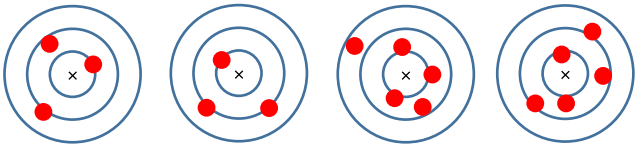
「偶然による誤差」は、測定のばらつきとして発見することができ、統計的な処理で誤差の大きさを推定したり、多数回測定を行い平均することで補正したりできます。

しかし、「系統的な誤差」は一方に偏ることになるため、平均しても取り除くことができません。「メスシリンダーの壁面に水滴がつくと、測定値が小さくなる」などのように観察で気づくことが出来るケースもありますが、気づかないことも多く、測定者の経験や注意深さ、実験計画の慎重さが必要です。



測定値の分布（統計学用語では「標本」という）を見るとき、ばらつきと偏りを考慮できるようになろう。
・ばらつき … 統計的に処理（判定）が可能。
・偏り（bias）… 原因が分かっていなければ、測定後の補正は難しい。実験計画の段階で考えておくことが大切。

2 真値の推定には、ばらつきの理論的分布を知る必要がある



ここでは、偶然誤差だけによってデータ（測定値）がばらつくものとして、真値を推定することを、考えよう。偶然によって起こるばらつきは、理屈では真値の両側に均等に広がると考えられます。しかし、上図のようにデータ数が少ないとき、データは真値の周囲に均等に広がるわけではなく、その平均が真値と等しいという保証はありません。つまり、平均によってばらつきを補正したとしても、ある程度の偶然（推定値が真値とずれること）は存在し続けると考えるべきなのです。定量的な科学において、扱う数値に偶然が含まれるのであれば、その偶然の大きさを定量的に示さなければ信頼性を判断できません。この「偶然を定量的に扱う学問」が、これから勉強する推測統計学です。

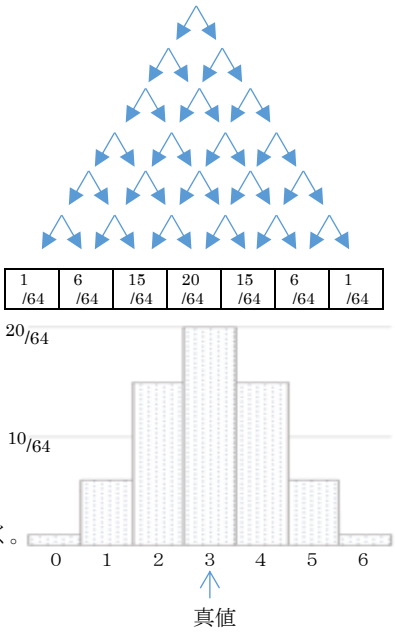
3 「正規分布」は、推測統計学において最も重要な理論的分布である

真値を知り得ない以上、測定値を眺めているだけでは、偶然を数値化することは不可能です。そこで、真値が分かっているものとして、偶然が生み出すばらつきの分布を理論的（数学的）に導くことにします。この分布は、19 世紀の数学者ガウスが導いたもので、正規分布（normal distribution）と呼ばれる確率分布です。（確率分布とは、確率変数Ｘ（データに相当）とそれに対応する確率 P(X)の関係（関数）を言います。今は、なんのこっちゃ？でＯＫ。そして、確率変数が離散型か連続型かで、確率分布は大別されます。正規分布は連続型確率変数をとる分布です。）

高１の皆さんにとって、いきなり連続型確率変数をとる正規分布の説明はハードルが高い（実は、連続や無限の概念が難しい）ので、離散型の確率変数をとる 二項分布（binomial distribution）の説明から始めます。

※「確率分布」の説明で、確率変数、確率、確率密度関数などの言葉に混乱する人が多いのですが、次ページの説明を読んで、頭を整理してください。

4　二項分布と正規分布



真値から偶然によってばらつきが生じることをイメージするため、パチンコ玉が釘にぶつかって左右に飛ばされながら、下に落ちていくことを考えます（右図）。パチンコ玉が左に行くか右に行くかは偶然で決まり、その確率は共に１/２であるとします。計算によって、６段の釘を落ちていくパチンコ玉がどの位置に達するかの確率（相対度数と考えてもＯＫ）を求めると、右の表やヒストグラムが得られます。このような確率分布を「二項分布」（binomial distribution）と言います。

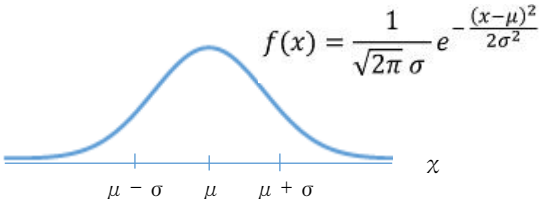
ここでヒストグラムについての約束を確認しておきます。ヒストグラムでは、各柱の面積がその階級値における度数を示します。上の話では、相対度数を求めていましたので、全ての柱の面積（相対度数）の合計は１になります。

＜今は読むな＞
今回の二項分布は、試行回数 n＝6、成功確率 p＝1/2 の二項分布で、これをＢ(6,1/2) と書く。そして、成功回数ｋに対する確率関数は、Ｐ(ｋ)＝$_6C_k$ (1/2) ^{k}・(1－1/2) $^{n-k}$ となる。
ｎやｐは予め与えられており、パラメータと呼ばれます。また、成功回数ｋは確率変数と呼ばれ、確率Ｐ(ｋ)に対応します。一回の試行における成功確率ｐがあったり、全試行の結果に対する確率Ｐがあったり、ややこしいですね。

「確率分布」と「相対度数分布」

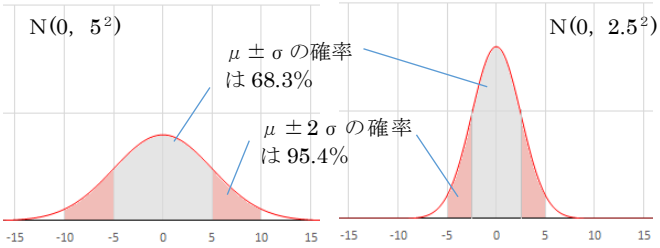
上の話は思考実験であり、得られた結果（パチンコ玉が達した位置Ｘとその確率Ｐ(Ｘ)の関係）は確率的（数学的）に導いたものです。これを「確率分布」 と言います。つまり、「確率分布」とは、理屈で導いた「相対度数分布」のことで、関数として示されるものを呼びます。関数では視覚化のためにグラフを用い、それは曲線ｆ(Ｘ)で示されることが多いのですが、確率分布Ｐ(Ｘ)のグラフでは、曲線で囲まれる面積がＰ(Ｘ)を示しています。曲線ｆ(Ｘ)には確率密度関数 という名前が付いています。その数式は難しいものが多いですが、統計ユーザーはＰ(Ｘ)を扱うことが出来れば十分です。多くの確率分布で、その確率Ｐ(ａ＜Ｘ＜ｂ)は表（「正規分布表」で検索してみてください）の形になって提供されています。

話を、偶然が生み出すばらつきの理論的分布（確率分布）に戻します。上のようなパチンコ玉の実験で釘の段数を無限に増やしていくと、確率分布のグラフは滑らかになります。これが「正規分布」（normal distribution）です。



正規分布は、連続型の確率変数 Ｘ をとります。その確率密度関数ｆ(χ) は右式で示されますが、上の囲みで書いたように、統計ユーザーは、この式の意味が分からなくても大丈夫です。しかし、ここでは少し頑張って、この式には、確率変数χの他に、平均μと標準偏差σ（もしくは分散σ^2）が入っていることを見ておきましょう。つまり、正規分布の形は平均μと分散σ^{2}で決まることを意味しており、正規分布は Ｎ(μ， σ^2)と表します。正規分布のμは、左右対称性から分布の中央に位置します。また、標準偏差σは、分布のばらつきを示すもので、σが大きくなれば分布の形は横長で背が低くなります。(ｆ(χ)の中のπ＝3.14…、ｅ＝2.71…は定数です。μやσ^{2}はパラメータと呼ばれます。慣れ親しんだ一次関数ｙ＝ａχ＋ｂのａやｂもパラメータです。グラフの形は、パラメータで決まります。)

最後に、正規分布の確率Ｐ(Ｘ)について、重要な性質を説明します。どのような正規分布でも、区間（μ－σ，μ＋σ）の面積（確率）は全面積の約 68.3%、区間（μ－2σ，μ＋2σ）の面積（確率）は全面積の約 95.4%という風に、平均μから標準偏差σの何倍まで離れた区間をとるかで、その確率が決まってきます。



2019 課題研究基礎　統計学② 事前プリント（宿題）　**※ 枠内に答えて、25 日(水)に職員室前 BOX に提出。**

1 年 1 組（　　）番　氏名（　　）

□ アンケート

Q1　誤差には、「偶然による誤差」と「系統的な誤差」があることを理解していますか？

①　前からよく理解している　　②　資料を読んで理解することができた

③　資料を読んだが、理解できていなかった　　④　資料を読んでおらず、理解できていない

Q2　「確率分布とは、確率変数と確率の対応を示す関数である」ということを理解していますか？

①　前からよく理解している　　②　資料を読んで理解することができた

③　資料を読んだが、理解できていなかった　　④　資料を読おらず、理解できていない

Q3　「確率分布のグラフで、確率は面積で示される」ことを理解していますか？

①　前からよく理解している　　②　資料を読んで理解することができた

③　資料を読んだが、理解できていなかった　　④　資料を読おらず、理解できていない

□ 問題

Q4　測定値の分布における「ばらつき」や「偏り」を補正し真値を推定する方法についての説明として、正しいものを 1 つ選べ。

①　ばらつきの補正には平均が有効だが、偏りの補正に平均は役に立たない。

②　偏りの補正には平均が有効だが、ばらつきの補正に平均は役に立たない。

③　偏りの補正にも、ばらつきの補正にも、平均は同様に有効である。

Q5　正規分布に関する説明として、正しいものを 1 つ選べ。

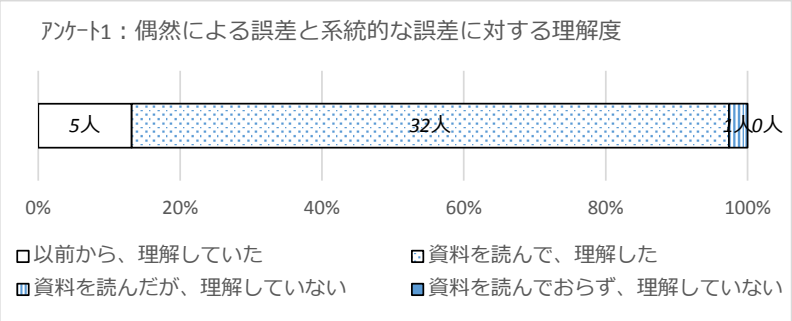
①　滑らかで左右対称な形で、その中央となるＸが平均である

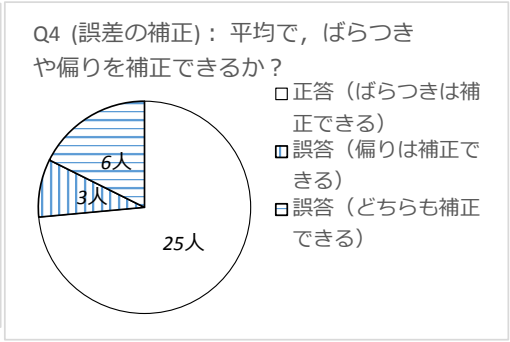
②　コイン投げで表の出る回数とその確率は、正規分布となる。

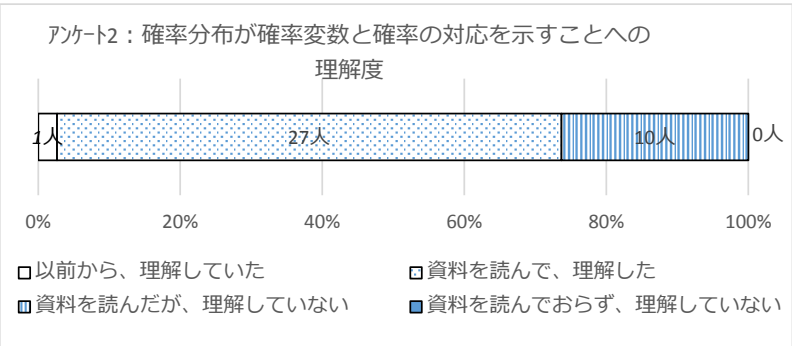
③　平均が同じであれば、正規分布の形は 1 つに決まる。

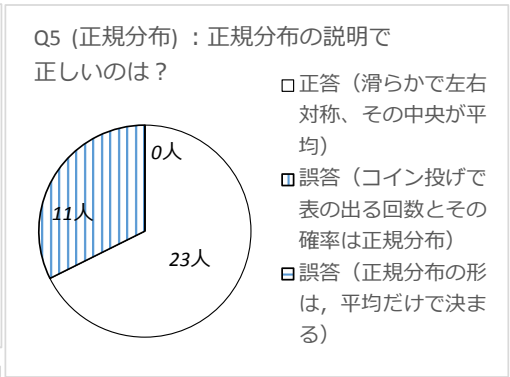
Q6　10000 人が受験したある全国模試の平均点が 50 点，標準偏差が 10 点であった。この模試の点数の分布を平均＝50，標準偏差＝10 の正規分布 N（50，10^2）であると考えたとき，模試の得点が 70 点以上の人の人数は何人と推定されるか。

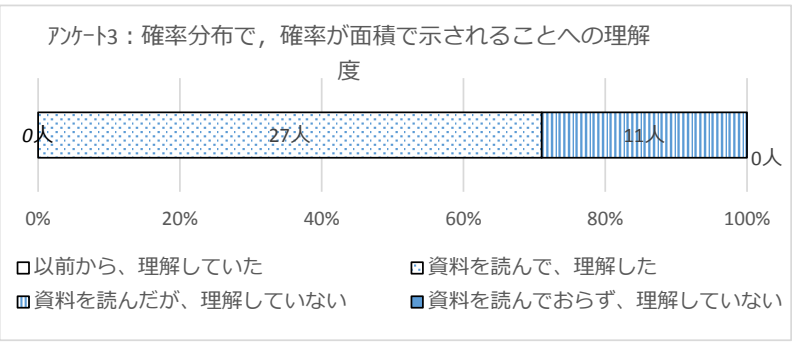
191008課題研究基礎(統計②)　事前学習アンケート　集計結果

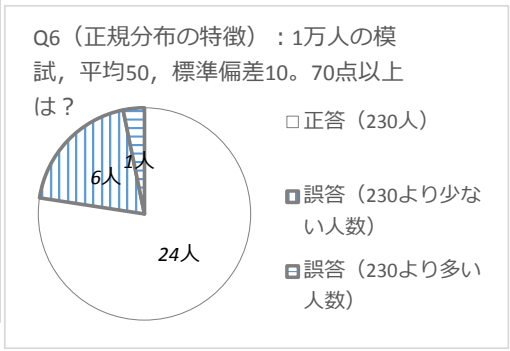






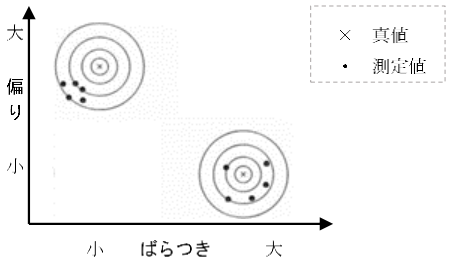






課題研究基礎　統計②　宿題の答え

Q4　①　ばらつきは、平均をとることで、打ち消しあう。



Q5　①

① 正規分布は連続型確率変数をとり、平均を中心とした左右対称な分布である。

② コイン投げについての分布は二項分布であり、離散型確率変数（表（おもて）の回数）をとる。

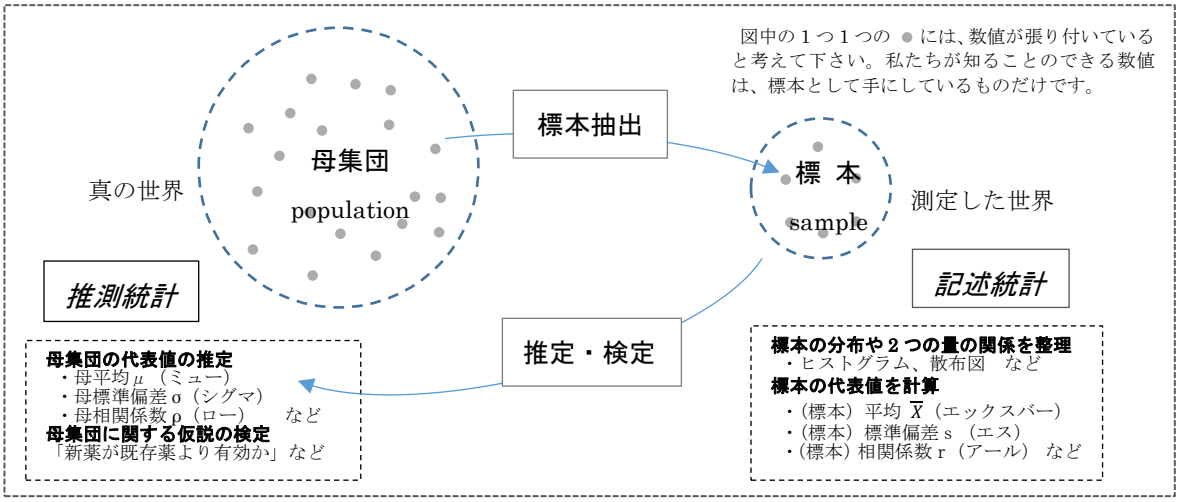
③ 正規分布の形は平均と分散が決まれば、1つに決まる。

Q6　230 人　平均＝50，標準偏差＝10 のとき，得点 70 は平均＋2×標準偏差である。正規分布では，区間（平均－2×標準偏差，平均＋2×標準偏差）の確率が約 95.4%であることが知られているから，平均＋2×標準偏差以上となる確率は，(100－95.4)／2＝2.3%である。よって求める人数は，10000×0.023＝230。

2019 課題研究基礎 統計学②（推測統計～母平均の推定～） 事前プリント その２

１．推測統計学

測定値をグラフにしたり平均をとったりする統計学を記述統計学と言いますが，次回の授業で扱う統計学は推測統計学と呼ばれています。研究者は、測定したデータの向こうに我々には知りえない真の世界（近づくことは出来る）があると考えているはずです。この真の世界を推測する手法が、推測統計学です。推測統計には「推定」と「検定」の２つがありますが，ここでは「推定」の基礎の基礎を学びます。推測統計では，数学者が理論的に（確率計算で）導いた確率分布を用います。学ぶべき確率分布はいくつかありますが，ここでは正規分布と t 分布を学習します。



２ 「偶然による誤差」を考えるための道筋 ― 前回の事前プリントの復習から始めます

標本の大きさ n が小さいとき

誤差：真値と推定値の差

測定の「偶然による誤差」を考えたい ― but → 測定値から正確な真値は知り得ないのだから、正確に誤差は判断できない

＜逆転の発想＞ 測定値から考えるのではなく、

真値 μ をわかっているものとして、偶然によるばらつきがどのようになるのかを、理論的に導こう。

＜数学の力を使って＞

理論的には、偶然による誤差は「正規分布」と呼ぶ確率分布で表される。

※ 正規分布 $N(\mu, \sigma^2)$ の形は、平均 μ と分散 σ^2 で決まる。

＜例のジレンマ＞ 測定値（データ）から、正確な μ や σ^2 を知り得ることは出来ない。つまり、正規分布の形を決められない。

しかし、手掛かりは測定値しかない。

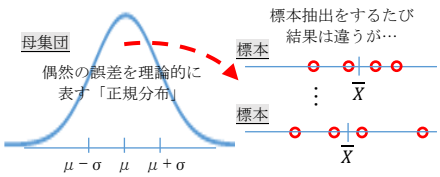
＜今回の事前プリントの内容＞ 複数の測定値から得る平均 $\overline{x}$ ・分散 s^2 と誤差のばらつきを示す「正規分布」の平均 μ ・分散 σ^2 はどのような関係にあるのか？

標本　標本平均・標本分散　母平均・母分散

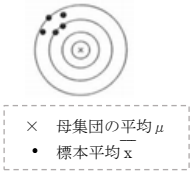
３ 母平均・母分散の推定値 ～不偏推定値～ としての標本平均・標本分散

１）標本平均

「実験でデータをとる」ことを，「正規分布から無作為に標本抽出する」ことと考えてみよう。標本から得られる標本平均 $\overline{X}$ は標本抽出するたびに違いますが，我々は標本平均 $\overline{X}$ を母平均 μ の推定値として（とりあえず）使っています。この意味を考えていくことにします。仮に標本抽出を何度も繰り返し，標本平均の分布をとったらどうなるでしょうか。直観的に標本平均の平均 $E(\overline{x})$ が μ と等しい正規分布（平均と分散で決まる）になるように思いませんか。分散は４で議論することとして，ここでは $E(\overline{x}) = \mu$ となることについて考えます。



$E(\overline{x}) \neq \mu$ のとき，$\overline{x}$ の分布は μ に対して「偏り」を持ちます（右図）。つまり，$E(\overline{x}) = \mu$ とは，$\overline{x}$ の分布は μ に対して「ばらつき」を持ってもよいが「偏り」を持たないことを意味します。このような推定値を不偏推定値と呼びます。$\overline{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$ が不偏推定値であることは，数学的に確認されています。次回の授業では，乱数実験で確かめます。



２）標本分散

まず，２）で議論する「標本ｘの分散」は，４で議論する「標本平均$\overline{x}$ の分散」とは違うので，混乱しないようにしてください。

標本分散には，標本の大きさ（データ数ｎ）で割る方法とｎ－１で割る方法の２つがあります。母分散σ^2の推定値として使われるのは，$s^2=\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1}$ で，数学的に証明されています。これを不偏分散と呼び，s^2ではなくu^2とも表記されます。不偏分散は，直観的に理解することが難しいと思います。次回の授業では，乱数実験で確かめます。

最後に，不偏推定値はあくまで「偏り」がないことが数学的に証明されているだけであり，１回の標本抽出で得られた推定値を信じて良いわけではないということを確認しておきます。得られた「標本平均」がどの程度「母平均」と一致するのかは，４で議論します。

【混乱しやすい統計用語】
①標本: 1セットのデータの集合。 ②標本の大きさ: 1つの標本を構成するデータの数。 ③標本の数: 標本がいくつあるかということ。

４　母平均の区間推定　―　母集団のσ^2が既知の場合

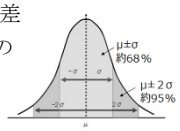
母集団について，我々が最も知りたいことの一つは母平均（誤差の議論では母平均μが「真値」を意味しています）です。ここでは，標本平均（実験データの平均）から母平均を考えるために，最も大切な定理を学びます。いきなりですが，例題を使って説明します。

定理１：　ｘが平均μ,分散σ^2の正規分布に従うならば、大きさ(データ数)ｎの標本抽出を何度も繰り返して得る標本平均 $\overline{X}$ の分布は，平均μ、分散σ^2/nの正規分布に従う。　→　授業では，乱数実験で確かめます。

※ この標準偏差 $\sigma/\sqrt{n}$ を「標準誤差」といいます

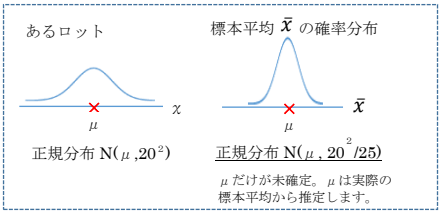
＜例題＞　ある薬（錠剤）の製造工場で、錠剤の重さが基準通りであるかを、管理している。この工場では、過去の経験から、１つのロット（同じ条件のもとに製造する製品の、生産・出荷の単位）での錠剤の重さはほぼ正規分布に従い、その標準偏差はσ ＝ 20（mg）であることが分かっている。管理者は、あるロットの中から錠剤 25 個を無作為抽出し、その重さの平均が $\bar{x}$ = 260（mg）であると求めた。ロット全体の錠剤の重さの平均μを信頼度 95%で区間推定せよ。

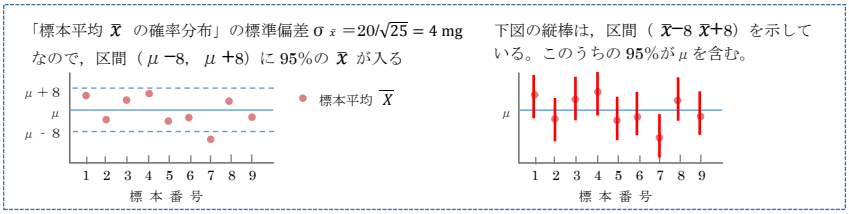
ただし，正規分布において区間（$\mu-2\sigma$, $\mu+2\sigma$）の区間確率は 0.95 であるとする。



＜解答＞　252～268 (mg)

＜解説＞① 問題中の数値を使って定理を図示。　②正規分布において区間（$\mu-2\sigma$, $\mu+2\sigma$）の区間確率が 0.95 であるから…





②で考えた通り、標本平均±8 をたくさん考えたときに、それらの 95%で母平均μを含むはずである。実際の標本抽出は 1 回だけで，その標本平均 $\bar{x}$ = 260（mg）であった。つまり，母平均μが 260±8（mg）の範囲に入る確率は，95%と推定される。

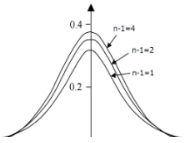
□ このように，推測統計では，物事を確定的に定めることはなく，確率的に定めます。「偶然による誤差」の議論を復習し，推測統計の基本的な考え方を整理しておきます。

無数の測定値を得れば，その分布は真値を平均とした正規分布(母平均μ，母分散σ^2)となるはずだが，母平均（真値）はわからない。そこで，データ数（標本の大きさ）ｎの標本抽出を行って標本平均$\overline{X}$ を求める。定理１によれば，無数の標本平均を得れば，その分布は正規分布(平均μ、分散σ^2/n)となる。つまり，標本平均は，母平均を中心にしてその両側に、標準偏差 $\sigma/\sqrt{n}$ （これを標準誤差という）でばらつく分布をとっている。正規分布では，確率変数の区間を決めれば，そこに入る確率（面積）が決まるので，これにより，母平均（真値）を確率的に推定することが可能となる。

５　ｔ分布を使う推測統計　―　母集団のσ^2が未知の場合

４の例題は，“母分散が既知のときに，母平均を推定する”ものでしたが，母分散が既知な場合は少ないはずです。では，必要な母分散σ^2が未知の場合，母分散の代わりに標本から計算した不偏分散s^2を用いて推定することは許されるでしょうか。答えは NO です。標本サイズが十分大きい場合を除けば，標本から計算される不偏分散s^2（不偏標準偏差ｓ）は，母分散σ^2（母標準偏差σ）との誤差が大きいと考えられるからです。　→ 授業では，乱数実験で確認します

そこで，σをｓで置き換える新たな方法として，数学者はｔ値という新しい変数 $t=\frac{\overline{X}-\mu}{s}\sqrt{n}$ を定義し，正規母集団から大きさｎの標本を取り出し，それからｔ値を計算するという抽出実験を繰り返してｔの度数分布をつくることを考えました。これを理論的に導いたのが右図のようなｔ分布です。標本の大きさｎによって形が異なり，ｎが十分に大きいと正規分布に重なります。ｔ分布を使うと，ｔ値に対する区間確率が分かるので，母平均 $\mu=\overline{X}-t\frac{s}{\sqrt{n}}$ を確率的に推定することが可能となります。



t分布：標準正規分布に似ている。標本数ｎから1を引いた、自由度＝ｎ－１で形が決まる。自由度が大きくなると、背が高くなり次第に標準正規分布に近づく。

2019 課題研究基礎　統計学② 事前プリントその２（宿題）　**※ 枠内に答えて、７日(月)に職員室前 BOX に提出。**

１年１組（　　）番　氏名（　　）

□ アンケート

Q1　「記述統計」と「推測統計」の違いを理解していますか？

①　前からよく理解している　　②　資料を読んで理解することができた

③　資料を読んだが、理解できていなかった　　④　資料を読んでおらず、理解できていない

Q2　統計用語の「標本」とは何か，理解していますか？

①　前からよく理解している　　②　資料を読んで理解することができた

③　資料を読んだが、理解できていなかった　　④　資料を読おらず、理解できていない

Q3　「不偏推定値」とは何か，理解していますか？

①　前からよく理解している　　②　資料を読んで理解することができた

③　資料を読んだが、理解できていなかった　　④　資料を読おらず、理解できていない

Q4　「推測統計では、確率分布(理論的に導いた分布。例えば、正規分布)の性質を利用して、
推定したい母集団の特性値(例えば、母平均)を測定データから推測する」ことを理解していますか？

①　前からよく理解している　　②　資料を読んで理解することができた

③　資料を読んだが、理解できていなかった　　④　資料を読おらず、理解できていない

□ 問題

Q5　「標本」に関する次の文を読み，正しいものをすべて選べ。

①　“標本数”が多いとは、標本に含まれるデータ数が多いという意味である。

②　“標本サイズ”が大きいとは、測定値が大きいという意味である。

③　“標本サイズ”が大きいとは、沢山の標本があるという意味である。　　注）標本サイズ＝標本の大きさ

④　“標本平均”は、沢山の標本の合計を標本数で除して求められる。

⑤　①～④に正しい説明はない。

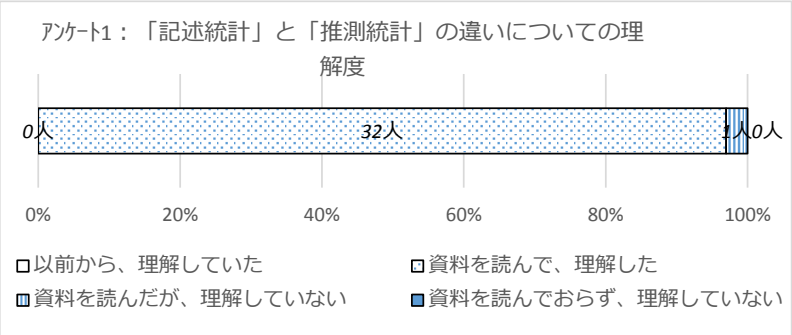
Q6　次の標本から，標本分散（不偏分散）を求めよ。

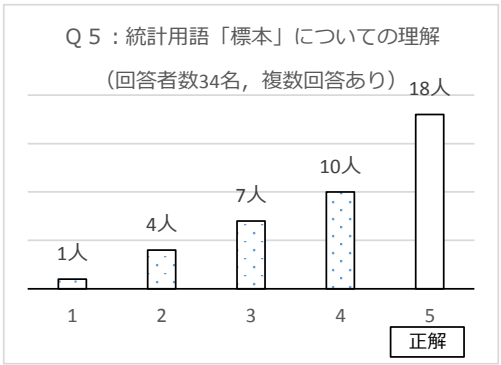
標本〔４０，５０，６０〕

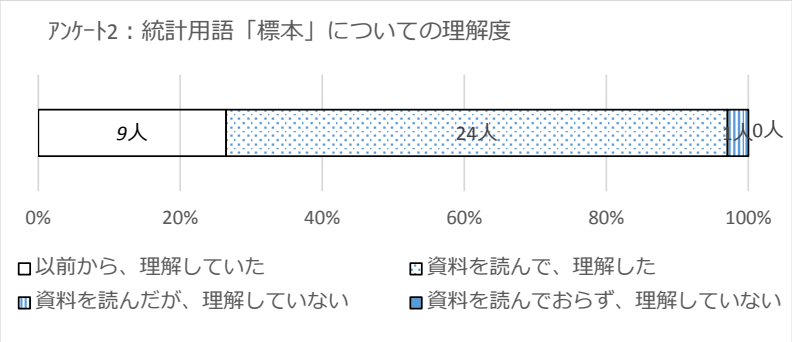
Q7　平均がμ，分散が 1 である正規分布から 100 個の標本を抽出したところ，
標本平均が３であった。平均μを信頼度 95%で区間推定せよ。

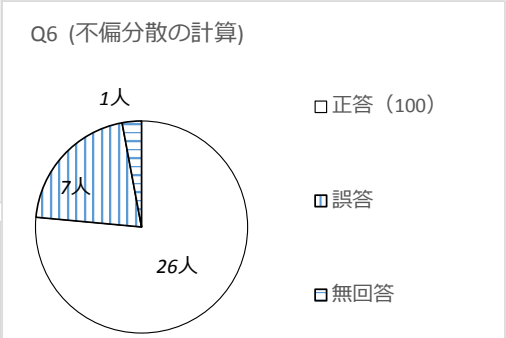
ただし，正規分布において区間（$\mu - 2\sigma, \mu + 2\sigma$）の区間確率は 0.95 であるとする。

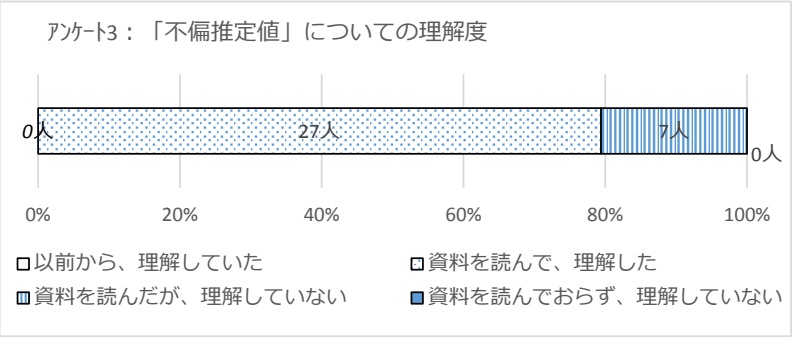
191008課題研究基礎(統計②)　事前学習 その２　アンケート　集計結果

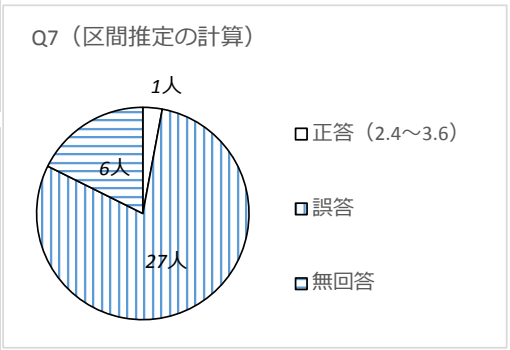


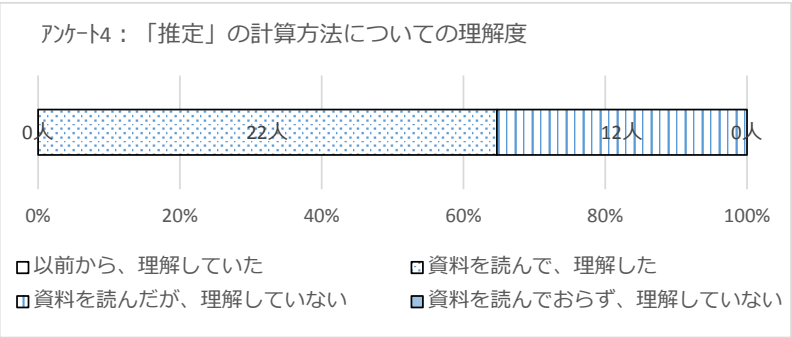












課題研究基礎　統計②その２　宿題の答え

Q5 ⑤　「標本」とは，1セットのデータの集合のことである。

- 「標本数が多い」標本が何セットもあるという意味。→①は間違い
- 「標本サイズが大きい」とは，標本に含まれるデータ数が多いという意味。→②と③は間違い
- 「標本平均」とは，標本に含まれるデータ（$\chi_1, \chi_2 \cdots \chi_n$）の合計を標本サイズｎ（データ数）で除したものである。→④は間違い

Q6 100　標本平均＝50，標本の大きさ＝3。不偏分散＝〔$(40-50)^2+(50-50)^2+(60-50)^2$〕／（3−1）=100

Q7 2.4～3.6　定理１より「標本平均 $\bar{x}$ の確率分布」の標準偏差 $\sigma_{\bar{x}} = 3/\sqrt{100} = 0.3$。この分布は正規分布であり，正規分布の性質から９５%区間は　$\bar{x} \pm 0.6 = 3 \pm 0.6$　となる。