高校数学
『偏差値』発　賢者の統計・愚者の統計 ようこそ！ 1

ー感動！ガリレオ流 帝都大学への数学　vol.7ー

2012 年 新課程より

『統計』のはしりは『数学Ⅰ：データの分析』として事実上誰もの必修科目へ！

今まで『確率・統計』を学ばなかった文系諸君が社会に出た後、数学の中で最も勉強の必要に迫られた唯一の単元は『統計』であっただろうことは容易に想像できます。

『統計』の一部内容が数学Ⅰにシフトした理由もここにありますが、はたして？？？

==========

発行日：2013 年 10 月 25 日

著者：ニ三五　八十三（フミコ　ハトミ）

発行元：「帝都大学へのビジョン」事務局

http://teito-vision.sunnyday.jp

teito-vision@ts.sunnyday.jp

==========

はじめに

本講座の意図は3つあります。

第1点目は、文字通り今の君たち（保護者さんの場合はお子さん）に一番縁が深い『偏差値』について、誤解のない理解をしておいてもらいたいという願いです。

実は、私たち日本人には『偏差値』という言葉は定着していますが、これを ”deviation value” や ”Z-score” や ”standard score” と訳しても英語圏の一般の方にはさっぱりと通じません。

統計を使っている人や専門家なら、“standard score”という英語を理解はしますが、日本の『偏差値』の意味とは少し違って捉えられるのです。

何故なら『標準偏差』などの世界共通の公式な統計用語とは異なり、日本で『偏差値』なる概念を教育に生かすようになってから初めてこれを説明するために英語表現ができたというような経緯がある言葉だからです。

マルチリンガルの松平先生（『ユダヤ式記憶術』などの著作者）にお聞きしたことなのですが、面白いことに『偏差値』をそのまま中国語読みすると中国人は理解するそうで、しかも、日本人と同じ意味に捉えるそうです。

さすが、「科挙」の国ですかね。
また、韓国でも理解はされると思いますが、中国・韓国いずれの国も『偏差値』システムが一般的に使われていることはありません。

ともかくも、自分に関わる身近なデータ『偏差値』の正確な意味を知らずに、世間の使い方から**何となく理解している諸君や保護者さんが多いのではないでしょうか？**

【保護者さんも必読！】という意味も込めて、「数学」と言えるほど難しいものではありませんから、なおさら、世間の間違った物言いだけで一喜一憂することのないように、その意味を正しく理解し、正しく参考にしていただきたいという願いをもって記しました。

1時間目：【偏差値】を求める・・・偏差という見方

1)　【学力偏差値】を日本語でほぐす

本講座のスタートは、君たちが受験の際に参考にする【学力偏差値】というものを、日本語で表現した計算式で説明するところから始めます。
尚、以降は、【学力偏差値】を単に【偏差値（Z-score）】と記述することとします。

君は、ある模試を受けました。
しばらくすると、採点結果として各教科の素点や平均点や順位とともに【偏差値】を受け取ります。

その【偏差値】とは、

$$君の偏差値 = \frac{君の得点 - 平均点}{標準偏差} \times 10 + 50$$ ・・・(1)

によって計算されている値なのです。

各教科での【偏差値】を算出する場合は、君の得点は素点そのもので、平均点も普通に考える平均点そのもです。

5教科総合での【偏差値】を算出する場合は、君の得点は5教科の素点の総計を考え、当然、平均点は各人の5教科の総計の平均点で考えればいいわけです。

意外に簡単な式ですよね。
記号で書かれた式よりもはるかに身近に感じ取れるのではないでしょうか？

ここまでに何度も繰り返し言ってきたことですが、ここでも、
『どんな複雑に見える式でも、日本語で理解しようとするクセをつけることが大切』
だということを再認識してください。

さて、(1)の計算式では、「君の得点」は君だけのデータですが、「平均点」と「標準偏差」は、君も含めた模試受験者全員のデータから割り出されるデータだということを最初に区別しておいてほしいのです。

さらに次のステップとして、この式(1)を見たときに、
君の得点はすでに決まっているとして、
ⅰ）平均点が低いほど僕の【偏差値】は高くなるなぁ！
ⅱ）【標準偏差】の値が小さいほど僕の【偏差値】は高くなるなぁ！
という具合なことが思い浮かべば nice ですよ。

こういう言い換えができることこそが数学という学問なんですよ。

さて、【偏差値】を決める(1)の式を理解する上で、君が知らないから理解できないのは現段階では【標準偏差】という言葉だけだと思いますが如何ですか？

実は、高校数学の『確率・統計』では【偏差値】という概念は残念ながら出てきません。
教科書に登場するのは、【標準偏差】という概念なんです。
【標準偏差】こそが、『統計』を学習する上で世界共通の最も重要で且つ基本要素なんです。

ただ、教科書には出てこない【偏差値】ですが、入試問題では過去には何度か出題されていますから、発展的な知識として知っていて当然と考えられているわけです。
自分の最も身近な値ですから、ここからしっかり学習しておいてくださいね。

では、第１ステップとして【偏差値】のイメージを掴んでいただきましょう。
【標準偏差】という概念をまだ知らないのですが、このステップでは、その【標準偏差】をおぼろげにイメージできればしめたものという目的で進めますね。
まずは、気楽に軽く通過してください。

Ⅰ）さて、例えば、君の偏差値が 70 だったとします。
（一般的には難関医学部を狙える高い偏差値ですね！）

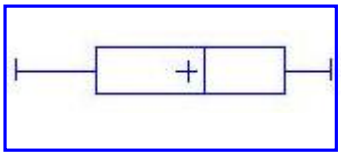
これを(1)式に当てはめてみましょう。
$\dfrac{\text{君の得点}-\text{平均点}}{\text{標準偏差}}=2$ であれば、偏差値がちょうど 70 になる計算になりますね。

そうです、
『偏差値 70』とは、『【標準偏差】の２倍分を平均点よりも上回った点数 get！』
ということだと言い換えられますね。

2時間目：【箱ひげ図】を書く・・・四分位という見方

1)　【箱ひげ図（box-and-wisker plot）】とは？

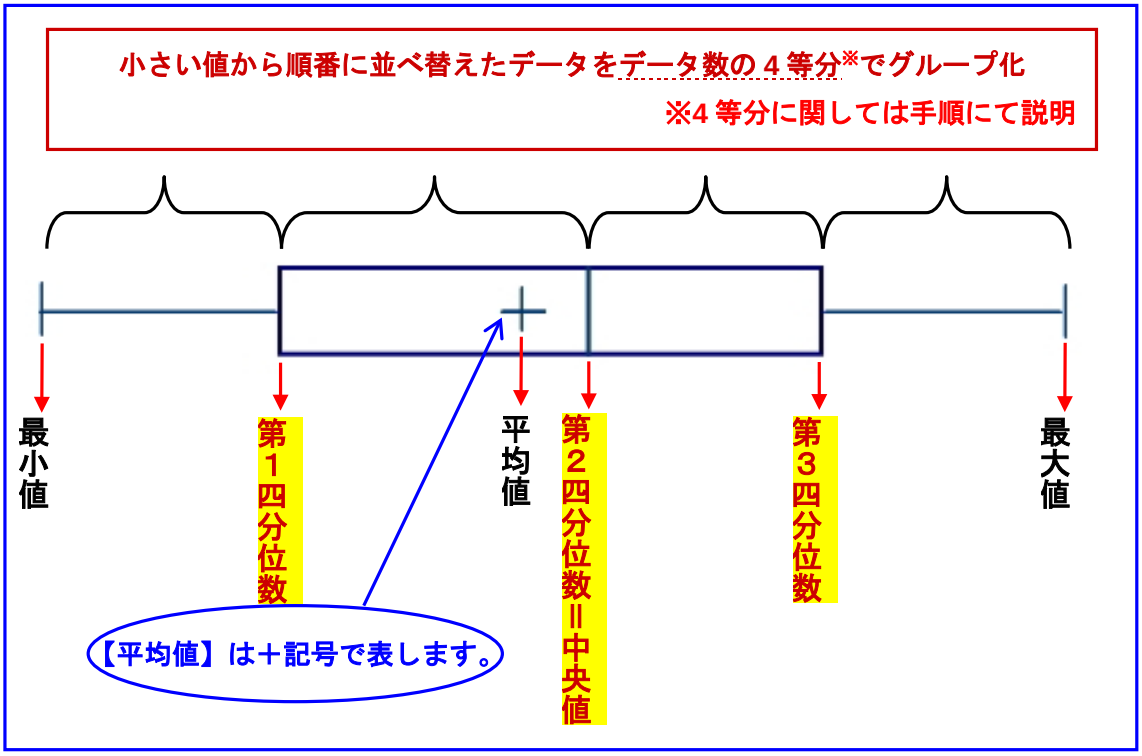
【箱ひげ図】は、全てのデータをデータ数が均等な4グループに分けて、その分け目の値3つと最小値、最大値、平均値の合計6個の値を一目見て分かるようにした図です。（右図）



ちょうど『箱』のような長方形と『ひげ』のような直線で表されることから【箱ひげ図（box-and-wisker plot）】と呼ばれています。

図の向きは、上図のように横向きに書く場合もあれば縦向きに書く場合もあります。
データの性質によって、フィットする向きに書けばいいだけの話なんです。
教科書では、たいてい横向きに書かれていると思いますので、本講座でも横向きで説明していきますね。

【箱ひげ図】を書く目的は、対象とする統計データの大よその分布具合をビジュアルに把握するためのものだということで今は認識しておいてください。先ず、【箱ひげ図】の意味と概念を下のビジュアルなイメージで捉えてください。



2)　【箱ひげ図】を書いてみる

この意味と概念さえ理解していれば、あとは、数学と言うより小学校の算数です。
簡単すぎて拍子抜けしてしまうかもしれませんよ。

『大よその分布具合』と言った理由が、この概念図で分かりましたか？
沢山あるデータの内、たった 6 個のデータで分布を表現しようと言うのですからね。

さて、知らない言葉が出てきました。
『第〇四分位数（しぶんいすう：quartile)』と『中央値（median)』なる 2 つの言葉です。
いずれも『統計』でしかお目にかからない言葉です。
少なくとも僕は仕事上でも今に至るまで、どちらの言葉も使ったことがありません。

でも、『四分位数』なる言葉は「小さい値から順番に並べ替えたデータをデータ数の 4 等分でグループ化」という言葉から、その由来も意味も想像がつくのではないでしょうか？

そこで、データを等しく 4 等分する手順に沿って説明して行きますが、具体的に分かりやすく理解してもらうために、サンプルデータで実際に作業しながら進めますね。
その元データは下表です。（10 名の国語のテスト結果）

生徒番号 I	1	2	3	4	5	6	7	8	9	10
国語得点 X	85	52	64	93	42	76	73	38	88	69

ⅰ）元データを小さい値から順番に並べ替えたデータを用意します。
サンプルの元データからですと、

38	42	52	64	69	73	76	85	88	93

となりますね。

ⅱ）次に、順番のままデータを 2 つのグループに等分します。
データ数は 10 個ですね。
ですから、「2 つのグループに等分する」ということは、小さい数 5 個と大きい数 5 個のデータにきれいに等分割できることになります。

38	42	52	64	69	73	76	85	88	93

3時間目：【正規分布】と【二項分布】・・・統計学の大黒柱

1)　【正規分布（normal distribution）】について

よく、

・偏差値が 60 なら自分より上位の人は全体の 16%
・偏差値が 70 なら自分より上位の人は全体の 2.3%
・偏差値が 80 なら自分より上位の人は全体の 0.2%

と思い込んでいる人が居ます。

実は、これが言えるのは『得点分布が【正規分布】である』という前提条件下なんですね。
これを知らずに、何でもかんでもこの物言いを当てはめていると賢い人の前では恥をかくことになります。

1 時間目に述べた本講座での国語(ex-1)の例は、正規分布に近い分布にしてありますので、上記のパーセントもかなり近似するはずですが、数学(ex-2)の例になりますと、正規分布とはかけ離れてきますので、もうこの解釈はあまり意味を持たなくなります。

ことのついでにお話ししておきましょう。
「学校の成績は正規分布しているものだ」と信じている諸君も多いのではないでしょうか？
世間では、何となく暗黙の了解でそのように考えてしまっている傾向があります。

しかし、現実的には、成績が正規分布している保証は全くありません。
どこの学校に行っても例外なく正規分布していたら逆に気持ち悪いですよね。

では、【正規分布（normal distribution）】について説明しておきましょう。
概念を説明するために、最初は仰々しい式をご紹介せねばなりませんが、この式自体は覚える必要などなく、後述する正規分布の持つ性質を理解しておくことこそが最も肝要なことです。

仰々しい式は正規分布の持つ性質を実現するための式なんだという認識さえあれば十分。
まずは、気楽に読み進めていってくださいね。

ここまで述べて来た偏差値のように離散的な階級、すなわち離散的な確率変数ではなく、連続的な確率変数が与えられ連続的に確率が分布すると考える場合、この分布曲線は、【確率密度関数（probability　density function）】と呼ばれます。

【確率分布関数（probability　distribution function）】と呼ばれそうなものですが、実は、2つの言葉は別物だということを理解しておいてください。

参考までに、【確率分布関数】の定義は、$(-\infty, x]$に収まる確率を表す関数なんです。
日本語でもう少し噛み砕いて言えば、$-\infty$からxに収まる確率のことを【確率分布関数】と定義するのですから、まさに『累積』のことになりますね。

ですから、【確率分布関数】は【累積分布関数（cumulative distribution function）】とも呼ばれます。

気が付きましたか？
言い換えれば、【確率分布関数】を微分したものが【確率密度関数】だということなんです。

高校数学では【確率密度関数】しか習いませんが、将来的に誤解することを避けるためにも、このことは教養として知っておいて損はありません。

さて、期待値（＝平均値）を$E(X)$、標準偏差を$\sigma(X)$として、【確率密度関数】が下記の式で与えられるとき、この分布を【正規分布（normal distribution）】と呼びます。

$$F(X)=\frac{1}{\sqrt{2\pi}\sigma(X)}\exp\left(-\frac{(X-E(X))^2}{2\{\sigma(X)\}^2}\right)=\frac{1}{\sqrt{2\pi}\sigma(X)}e^{-\frac{(X-E(X))^2}{2\{\sigma(X)\}^2}} \quad \cdots \text{(3-1)}$$

指数の部分がガウス関数であるであることから、別名【ガウス分布（Gaussian distribution）】とも呼ばれています。
先ほど申し上げたように、この仰々しい式を覚えておく必要はありませんよ。

式を見やすくするために標準偏差の(X)は省略し、ここからは統計学として語るために、期待値E　を平均値m　に置き換えておきます。

4時間目：相関というもの・・・【散布図】【共分散】【相関係数】

1) 相関関係を【散布図（scattergram）】で表す

2つの事象に何らかの相関関係があるかどうかは、いろんな分野いろんな場面で日常茶飯事に考えられているテーマなんですね。

それ故、その相関関係の具合を感覚やイメージでなく数値の程度で判断したいと望むのは、人間の理性として当然の行き先になるわけです。

この行き先を学んでもらうために、サンプルとして、10人の生徒の国語と数学のテスト結果データを3つ用意してみました。（3つ目は途中でご紹介します）
※注：学習用の創作データであって実際のデータではありません。

サンプルデータ1

生徒番号 I	1	2	3	4	5	6	7	8	9	10
国語得点 X	85	52	64	93	42	76	73	38	88	69
数学得点 Y	85	46	62	87	34	62	77	32	93	72

サンプルデータ2

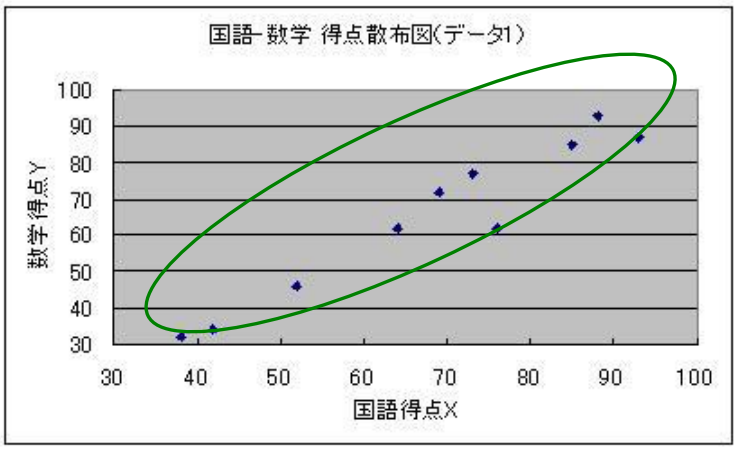
生徒番号 I	1	2	3	4	5	6	7	8	9	10
国語得点 X	85	52	64	93	42	76	73	38	88	69
数学得点 Y	72	93	32	77	62	34	87	62	46	85

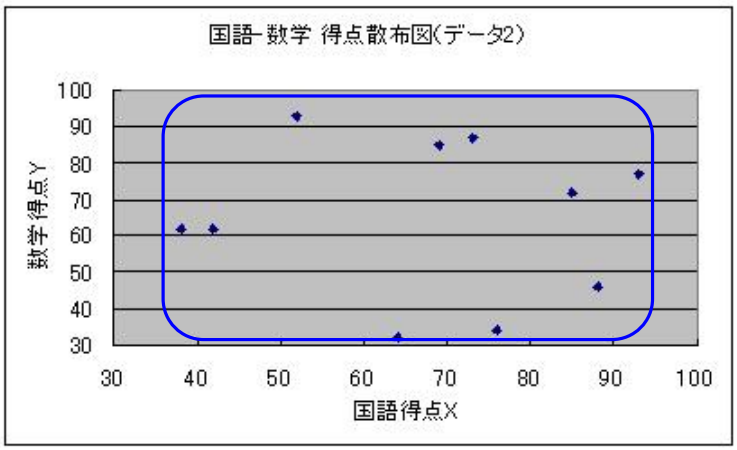
データ1は、僕が意図を持って作成したデータですが、データ2はデータ1において国語の成績をそのまま、数学の成績を単にデータ1と逆順に並べ替えただけの無意図のものです。

それぞれのデータで、国語と数学の成績には何らかの深い相関関係が存在するのだろうか？
誰しもが想定する問題ですよね。

そんなときは、先ずはグラフに表してみて、視覚的に感じることが第一歩なんですね。
グラフと言っても、今回は確率変数が2つあるから度数分布とはちょっと異なりますよ。

それでも、紙面は2次元使えるんだから、2つの変数を縦軸と横軸に利用すればいいのです。
横軸を国語の成績X、縦軸を数学の成績Yとして、それぞれのデータをプロットしてみた結果が次の【散布図（scattergram）】と呼ばれるグラフになります。



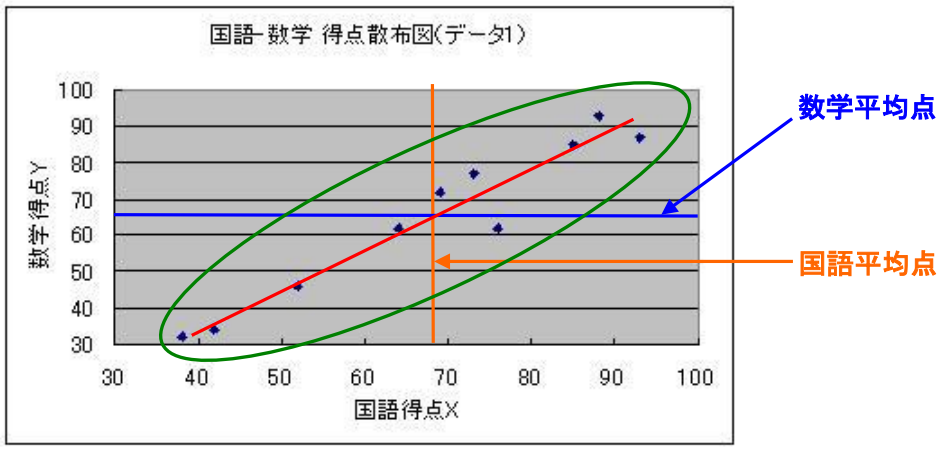


2つの【散布図】には明らかに違いがありますね。

データ１：国語の点が高ければ数学の点も高いという傾向が見えます。（　　）

データ２：数学の点は国語の点には全く関係なくバラついていると見えます。（　　）

さて、平均点は簡単に計算できますから、これをこの散布図に書き込んでおきましょうか。
データ１の散布図だけで示しておきますね。



5時間目：誤解しない【中心極限定理】・・・【推定】【検定】へ

1)　【推定】【検定】とは何をするのか？

僕が受験生だった41年前には、高校数学の【統計】には【推定と検定】というテーマが存在したんですけれども、平成に入ってからは長らく姿を消してしまっていたようですね。

とは言っても、大学に入学すれば、経済学や心理学はじめ文系諸学も、必ず勉強することになっているはずでしょうし、高校生でも塾や予備校で勉強している子はしているでしょう。

実は、実務社会に出てから実際に遭遇する可能性が最も高い単元の『統計』の中でも、経営や企画はもちろんのこと、実際に要求されるのが【相関関係の分析】とともに【推定と検定】なんですが、大学に入ってからは勉強しないもんだから、社会人になってから必要に迫られて勉強する方も結構多いのではないかと思います。

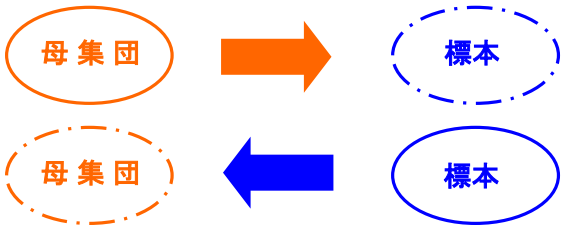
【推定と検定】なるテーマも、学ぶ内容だけを見ると、別段大した内容には見えないのですが、意外に分かりにくく深いものなんですね。

簡単な日本語なのに何のことを言っているのか理解できないと感じる方が多いと思いますし、はたまた、大いに勘違いをしたまま大人になられた方も多いと思われます。
それは、陥りやすい間違いを参考書では指摘されていないことにあるように見受けられます。

その誤解を生みやすい一つに【推定と検定】の根拠を提供する【中心極限定理】があります。

この【中心極限定理】は、母集団の【平均値】と【標準偏差】が決まっていることを前提として、母集団から抽出する標本の性質を数学的に規定するというものですが、現実問題としては、そもそも、母集団の【平均値】と【標準偏差】を知りたいがために標本を抽出して分析しますね。

【中心極限定理】は、確定した母集団から標本の性質を規定するもの。



現実の課題は、抽出した標本から母集団の性質を推定すること。

先ずは、この方向性の違いが誤解の根源になっているということを頭に入れてください。

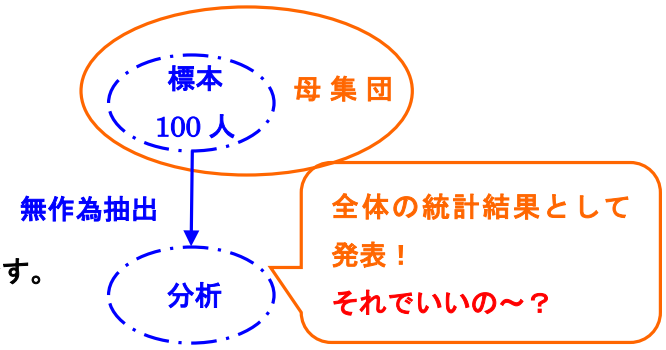
『【推定】【検定】とは何をするのか？』
これを最初に認識しておかないと、何をしているのかがサッパリわかりません。

本講座は、一貫して【偏差値】に関するサンプル素材を元に統計の基本を解説してきました。
1 時間目では、15,000 人の模試成績を【母集団（population）】として、ここから無作為に 100 人の成績を抽出し、これを【標本（sample）】として分析してきましたね。

そもそも、たった 100 人の成績を無作為に抽出して、それを 15,000 人の成績の統計として代表させていいのでしょうか？

母集団の平均値・標準偏差と標本の平均値・標準偏差は相違して当たり前ですものね。

このことこそが【推定】のテーマそのものです。



ここからは、母集団の【平均値】を M 、【標準偏差】を S と表し、【母平均】・【母標準偏差】と呼びます。

一方、標本の【平均値】は m 、【標準偏差】は σ と表して母集団のそれとは区別することにし、それぞれは、【標本平均】・【標本標準偏差】と呼びます。

また、一つの標本として無作為に抽出されたのデータ数を n 個と置きます。
この n を、『標本の大きさ』あるいは『標本サイズ』と呼びます。

注意！
『標本数』と呼ぶと誤解の元（後述）

【推定】とは、標本から得られた統計量【標本平均】・【標本標準偏差】から、求めたい母集団の【平均値】や【比率】が、一定の確率（信頼度）の下にどの範囲にあるかを絞り込むことを言います。

謂わば、ある信頼度の下で語り得る母集団の【平均値】や【比率】の区間を明らかにすることを言うのですね。
その意味で【区間推定】という言葉で補完しておくと頭の中の整理が進みます。

一方、【検定】とは、標本から得られた統計量から母集団や標本に関して特定の仮定（帰無仮説）を設定し，その仮定した統計量が得られる確率を求め、一定の確率（危険率＝有意水準）より低いか高いかで仮説の妥当性を判定することを言います。

謂わば、ある有意水準（危険率）の下で語り得る母集団や標本の統計量に関する仮説の真偽を明らかにすることを言うのですね。
その意味で【仮説検定】という言葉で補完しておくと頭の中の整理が進みます。

固い言葉では分かりにくいかもしれませんので、下に具体的な例を示しながらまとめておきますね。
実際の考え方は、この後で例題を通して一緒になぞっていきますので、今は【推定】と【検定】の相違を理解するだけに止めてもらって結構です。

① 【推定】とは、
標本の統計量 → 母集団の統計量（平均値・比率）を一定の信頼度の下に推察すること。

[例]
【母集団】15,000 人の模試成績を、標本として 100 人の成績を無作為に抽出して集計したところ、【標本平均値】$m=57.7(\text{点})$、【標本標準偏差】$\sigma=12.1(\text{点})$となった。
95%の信頼度で【母平均】として公表できる点数の範囲を求めよ。

② 【検定】とは、
標本の統計量 → 母集団あるいは標本に対する仮説を立て、その妥当性を一定の危険率（有意水準）の下に判定すること。

[例]
【母集団】15,000 人の模試成績を、標本として 100 人の成績を無作為に抽出して集計したところ、【標本平均値】$m=57.7(\text{点})$、【標本標準偏差】$\sigma=12.1(\text{点})$となった。
【母平均】は 55 点以下であると言えるだろうか？
有意水準 5%で検定せよ。

【推定】は標本から直接母集団を推測し、【検定】は標本の結果から仮説されたことが、妥当かどうかを判定するということです。
要するに、「結果を予測する方向」か「仮説された結果を判定する方向」かという立論の方向性が逆であるということができますね。