Sample

統計検定®で学ぶ！

高校生からの統計学

代ゼミライセンススクール
Yozemi License School

Sample

受講をお考えの皆様へ

人工知能‘AI’が脚光を浴びています。
チェス・囲碁・将棋の全てにおいて、世界チャンピオンや名人を打ち負かした事実は皆さんの記憶にも新しいことと思います。金融取引の現場はもとより、今やタクシーの配車計画や通販のコールセンター対応に至るまでAIの守備範囲は広がっています。現存する職種の半分以上が消滅するとまで言われていますが、そんな時代に何を学んでおくべきか、更にはいかに学ぶべきかは、私たちにとって喫緊の課題です。

しかし、人工知能はあくまで『人工』の知能ですから、人間の思考を模して作られているものに他なりません。ですから、必要以上に怖れることもないと考えます。そもそもAIは機械学習を基盤とします。機械学習には最適化学習、深層学習(ディープラーニング)、ライブラリ学習、ベイズ学習などのアプローチ手法がありますが、それらと大いなる重なりを持つのが確率・統計です。

その「統計」および「確率」を学ぶための講座として、「統計検定®で学ぶ！高校生からの統計学」は開発されました。現代社会を生き抜くには必須分野である統計と確率を、基礎からしっかり学ぶことができます。

本講座では、統計検定®３級の過去問約50題を中心に、演習問題が豊富に掲載されています。理解するのに少し困難な部分があるかも知れませんが、繰り返し復習してみてください。有意義だったと振り返ることが出来る時間を皆さんと共有したいと思っています。

この冊子は、本講座の雰囲気を少しでもお伝えするために作成いたしました。主に統計を学ぶための「本編」と、確率を学ぶための「オプション編」の２種類で本講座は構成され、本冊子は「本編」テキストの一部を掲載したものになります。実際のテキストでは２種類合計で３００ページ超、講師オリジナル問題や大学入試問題までも活用した演習問題を含んだ大ボリュームとなっています。

また、本講座の一部を体験動画として下記サイトにて視聴いただくことも可能です。問題を解くことにこだわった統計学講座、ぜひ受講してみてください。

「統計検定®で学ぶ！高校生からの統計学」ホームページはこちら
※講義の一部を体験動画として公開しています。
【URL】http://www.yozemi-eri.com/license/statistics/

はじめに

西岡 康夫

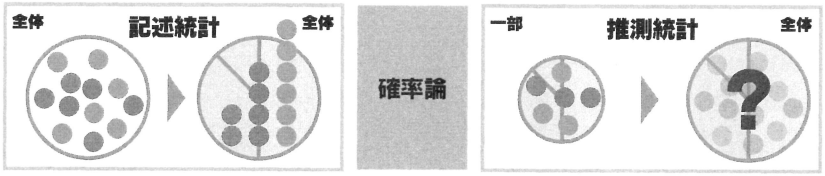
統計学は大きく２つに分類されています。**記述統計**と**推測統計**(統計的推測)です。

記述統計とは、考察の対象となるデータの観測されたすべての変量の平均、分散、偏差などを定量して、その分布を明示することで、データの特徴を把握する分野です。典型的な例としては、模擬試験における平均点や偏差値などが挙げられます。ざっくり言えば、「全体から部分へ」則ち、**総合 → 分析** という指向性をもっています。

それに対し推測統計は、考察の対象となるデータにおいて観測できた一部の変量(標本)から、データ全体(母集団)の特徴を**確率**(probability)という概念を用いて推論する分野です。典型的な例としては、選挙の出口調査やメディアの視聴率調査などが挙げられます。こちらもざっくり言えば、「部分から全体へ」則ち、**分析 → 総合** という指向性をもっています。

かつては大きなデータを作成するには物理的･経済的な限界がネックとなっていたので、推測統計が発達したという経緯があったのです。しかし、テクノロジーの進化に伴い、これまでは収集したり取り扱うことが容易でなかったビッグデータをかなり自在に入手できるようになり、スモールデータ(標本)を抽出することにそんなに拘らなくてもよくなりました。その結果、記述統計の果たす役割は大きくなってきています。

統計検定®３級の試験は、この記述統計を主範囲とし、併せて推測統計の一部を問いかけるものです。詳細な知識は求めていませんが、理解の幅を拡げておくことが合格には必要でしょう。



データ全体の特徴を要約する分野
[例.模試における平均点や偏差値]

一部のデータに基づいて全体の特徴を推測する分野
確率論を用いる
[例.選挙の出口調査、メディアの視聴率調査]

ただ、ビッグデータにも落とし穴があることを忘れてはいけません。

‘Garbage in garbage out’

ゴミのようなデータを大量に集めても、得られる結論はゴミという警句は統計学の世界でももちろん健在です。

テキストの構成と活用法

■テキスト構成

このテキストは統計検定®３級の過去問題を中心に編集されています。

これまでに統計学に関する授業等を受けたことがない方や、高校数学Ⅰ「データの分析」の内容に自信が無い方は、はじめに§０から受講されることをお勧めします。グラフや表を用いて用語を一つひとつ丁寧に解説しています。

§１以降は問題演習を通して統計学の知識を学んでいきます。過去の検定問題から統計学を学習するのに適した良問を選び、さらに、見開き頁の半分をノート用として紙面を割いています。設問を解いた際、なぜその選択肢を選んだかの根拠をメモしたり、講義における板書内容を書き写したりしていくことで、問題への理解を深めていくことができるでしょう。

■テキスト活用法

講義内で解説される授業内容を自らのものにするためにも、“予習”をされてから講義を視聴してください。上述のように根拠や疑問点をメモすることで、より深い理解が期待できます。

もしも理解が不十分と感じられた場合には、再受講制度も活用していただけますので、学習内容の定着や理解度アップにお役立て下さい。

目　　次

§0 データの分析

▶0.1.1 統計 (statistics)

統計とは，集団として存在する対象の属性を定量的に把握することを意味します。

▶0.1.2 対象 (object)

認識，感情，意志，想像などが働きかける‘もの’を一般に対象といいます。客体と同じ意味で用いられ，その実在の有無は問われません。

人物はもとより自動車とか飛行機といった無機物でも，実在の確認ができない霊魂などでも対象とよべるということです。

▶0.1.3 属性 (attribute, property)

対象に固有の性質や特徴のことです。狭義には，それが存在してはじめて対象を認知しうるような性質のことを属性といいます。

‘優しい’とか‘我慢強い’といった定性的属性，‘重さ’や‘長さ’といった定量的属性があります。例えばブドウという対象を考えるとき，‘甘い’は定性的な属性の表現ですが，‘糖度16度’は定量的な属性の表現となります。

統計において，集団とは人やものの集まりを指すだけでなく，1個人であっても時間変化に伴う数値(体重，血圧，血糖値 etc.)の集まりなども示します。要するに，同一の属性に関わる複数の数値の集まりを集団とよびます。

▶0.2.1　与件（data）

考察の対象を判断したり，他者と共有するために，その対象に定性的な形式化や定量的な数値化･符号化などを施したものを**与件･所与**などといいます。

▶0.2.2　資料（information）

'information' は情報と訳されることが多いのですが，こちらは「加工されていない生の資料」という意味が近いようです。

▶0.2.3　情報（intelligence）

資料に洞察を加え，問題解決や，意思決定に際し有用な形にしたものをとくに**情報**とよびます。

▶0.2.4　変量・変数（variable，variate，variant）

日数，日時，費用…　などのように，ある集合に含まれる要素(元)の性質を数値的に表現したものを**変量･変数**といいます。

▶0.2.5　尺度（scale）

変量と数値を対応させる基準のことです。
名義尺度，順序尺度，間隔尺度，比例尺度(比率尺度)があります。

▶0.3.1 分布 (distribution)

ある変量について，その測定値の広がりの状態，散らばり方を**分布**といいます。

▶0.3.2 代表値 (representative value, central value)

データの分布の状況を特徴的な唯一つの数値で表すとき，その値を代表値といいます。代表値には，平均値，中央値，最頻値があります。代表値は**要約統計量**(summary statistics)とか，**記述統計量**(descriptive statistics value)ともよばれます。

▶0.3.3 中央値 (median)

データをその値の大きさの順に並べたときに中央の位置にくる値を**中央値**といいます。データの個数が奇数の場合は，データの中に中央値になるものが存在しますが，偶数の場合は，中央の値が存在しません。そこで，もっとも中央に近い2個のデータの平均を中央値とします。

▶0.3.4 最頻値 (mode)

データの中でもっとも個数の多い値を，そのデータの**最頻値**といいます。一番売れ筋の商品が知りたい場合は，最頻値が有用です。最頻値は名義尺度以上で意味をもちます。

▶0.3.5 相加平均 (arithmetic mean)

算術平均ともいい，要素の数値総和を総個数で割った値を指します。
例えば，学級内のある5人の期末試験における国語の点数が72点，66点，87点，74点，91点であるとき，相加平均(平均点)は次のように求められます。

$$\frac{72+66+87+74+91}{5} \quad (=78)$$

▶0.3.6 相乗平均 (geometric mean)

幾何平均ともいい，要素の数値総乗の総個数累乗根をとった値を指します。
例えば，学級内のある5人の期末試験直前の国語の勉強時間が2分，8分，16分，16分，256分であるとき，相乗平均は次のように求められます。

$$\sqrt[5]{2\times 8\times 16\times 16\times 256} \quad (=16)$$

一方，この場合の相加平均を求めると，

$$\frac{2+8+16+16+256}{5} \quad (=59.6)$$

となります。

▶0.4.1 度数分布 (frequency distribution)

あるデータにおける変量を分類したリストを**度数分布**とよびます。データの中で最も大きな変量の値を**最大値**(maximum value)，最も小さな変量の値を**最小値**(minimum value)とよび，最大値と最小値の差を**範囲**(range)といいます。

範囲＝最大値－最小値

▶0.4.2 度数分布表 (frequency table)

一般に変量の大小の順(昇順・降順)で並べ，範囲を一定幅の小区間＝**階級**(class，**ビン**：binともいう)で区切り，その階級に存在する**度数**(frequency：変量の個数)で表現した表を**度数分布表**といいます。

また，その階級の幅を与える両端の値の平均(階級内の最大値と最小値の相加平均)を**階級値**(class value)といいます。

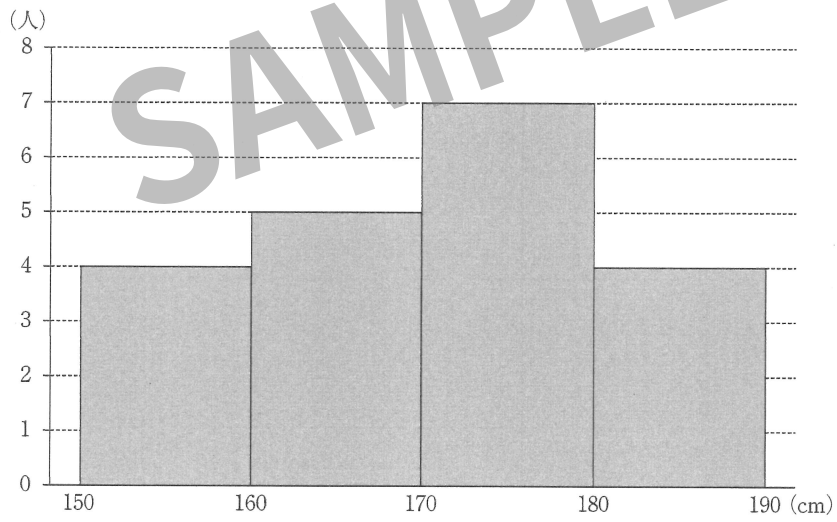
階　級	148以上～156未満	156以上～164未満	164以上～172未満	172以上～180未満
階級値	152	160	168	176
度　数	6	8	18	8

▶0.4.3　度数分布図・ヒストグラム（histogram）

度数分布をグラフとして再表現したものを**度数分布図**とかヒストグラムとよびます。

よく間違われることなのですが，ヒストグラムは単なる棒グラフではありません。棒状(柱状)のグラフではありますが，階級幅が一定であること，階級の度数がそのグラフの高さで表されることが不可欠です。また，各棒は離れません。

というのも，そのグラフにおける任意の区間が与える面積が，度数と一致するという特徴を失いたくないからで，連続変数の場合に用いられる分布のグラフが，横軸と曲線が囲む面積に意味を持つことと対応させる狙いがあります。



▶0.4.4　幹葉図（みきはず：stem and leaf plot）

鉄道やバスの発車時刻案内表示では，縦に並べられた各時に対応する分の値だけが横一列に羅列されますが，これと同じように，縦軸の数値に基づいた値を横軸方向に並べて表示する図を**幹葉図**といいます。

例えば，ある中学校のクラスにおける男子生徒の体重が，(kg：小数点以下は四捨五入)

55　57　68　64　48　58　58　52　61　77　65　59　58　63　72

であるとき，幹葉図は次のように表されます。

(幹)	(葉)	(度数)
40	8	1
50	2 5 7 8 8 8 9	7
60	1 3 4 5 8	5
70	2 7	2

▶0.4.5 相対度数分布表 (relative frequency table)

各階級の度数の全体の総度数に対する比率を**相対度数**(relative frequency)といいます。

階　級	148以上～156未満	156以上～164未満	164以上～172未満	172以上～180未満
階級値	152	160	168	176
相対度数	0.15	0.2	0.45	0.2

▶0.4.6 累積度数分布表 (cumulative frequency table)

はじめの階級からその階級までの度数の和を**累積度数**(cumulative frequency)といいます。

階　級	148以上～156未満	156以上～164未満	164以上～172未満	172以上～180未満
階級値	152	160	168	176
累積度数	6	14	32	40

▶0.4.7 累積相対度数分布表 (cumulative relative frequency table)

はじめの階級からその階級までの相対度数の和を**累積相対度数**(cumulative relative frequency)といいます。

階　級	148以上～156未満	156以上～164未満	164以上～172未満	172以上～180未満
階級値	152	160	168	176
累積相対度数	0.15	0.35	0.8	1.0

▶0.5.1 四分位数（quartile，hinge）

与えられたデータを昇順(小さい方から大きい方の順)に並べ，それらを4等分した個数で区切るとき，順に1，2，3の区切りの境界が現れますが，その境界に最も近い位置にあるデータの値を順に，第1四分位数，第2四分位数，第3四分位数とよびます。分位数は分位点ともよばれます。

中央値は第2四分位数のことです。データの個数が必ずしも4で割り切れるわけではないので，分位数(分位点)の決め方には種々の方法があります。次に一例をあげます。

ⅰ）データの個数nが偶数の場合

データを昇順に並べて小さい方の$\frac{n}{2}$個のデータを下位データ，大きい方の$\frac{n}{2}$個のデータを上位データとします。下位データの中央値を第1四分位数(下側四分位数)Q_1，上位データの中央値を第3四分位数(上側四分位数)Q_3とよびます。

ⅱ）データの個数nが奇数の場合

データを昇順に並べて，先ず中央の値を除外します。残りの$n-1$個のデータについて，小さい方の$\frac{n-1}{2}$個のデータを下位データ，大きい方の$\frac{n-1}{2}$個のデータを上位データとします。下位データの中央値を第1四分位数(下側四分位数)Q_1，上位データの中央値を第3四分位数(上側四分位数)Q_3とよびます。

第1四分位数Q_1と第3四分位数Q_3の差を**四分位範囲**(IQR：inter quartile range)'IQR'とよびます。

$$\mathrm{IQR}=Q_3-Q_1$$

この四分位範囲を2で割った数値$\frac{Q_3-Q_1}{2}$を，**四分位偏差**(quartile deviation)といいます。

▶0.5.2 5数要約（five-number summary）

四分位数における5つの代表値，

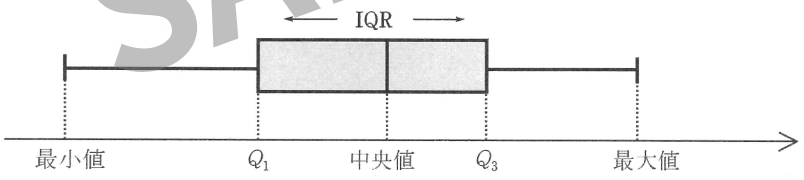
最小値，第1四分位数，中央値(第2四分位数)，第3四分位数，最大値

を「5数」とよび，これらを一覧にした表を**5数要約**といいます。次表は，ある中学校のクラスにおける女子生徒の体格を5数要約した例です。

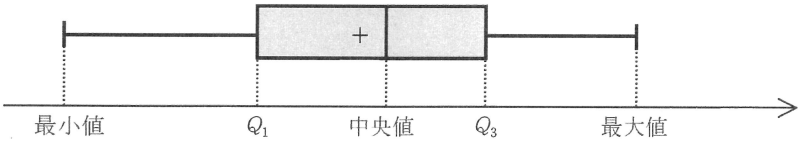
	身長(cm)	体重(kg)
最小値	148	39
第1四分位数	155	42
中央値	159	47
第3四分位数	162	56
最大値	173	74

▶0.5.3 箱ひげ図 (box and whisker plot)

データの分布，特に5数要約を可視化するものとして，**箱ひげ図**とよばれるグラフがあります。これは，データの最小値，Q_1，中央値(平均値の場合も)，Q_3，最大値を次図のように箱とひげで表したものです。



※　箱の長さは四分位範囲(IQR)を表します。



※　平均値を＋で表現し挿入する場合もあります。

▶0.5.4 外れ値 (outlier)

データに含まれる変量の中で，際立って大きな値や小さな値を示すものを**外れ値**とよびます。外れ値は平均値や分散などの統計量への影響が大きいので，その影響をできるだけ除く対応が求められます。

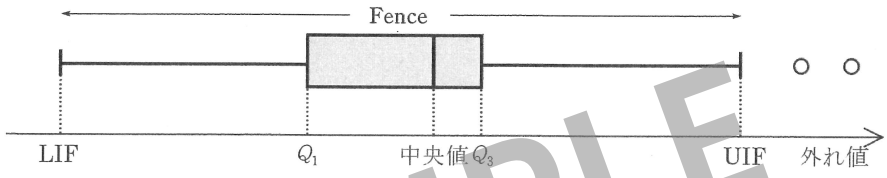
箱ひげ図の描き方においても，外れ値になりやすい最大値や最小値などを明示するために，ひげの上限を，UIF(upper inner fence)とよび，

$$\mathrm{UIF}=Q_3+1.5\times(\mathrm{IQR})$$

ひげの下限を，LIF(lower inner fence)とよび，

$$\mathrm{LIF}=Q_1-1.5\times(\mathrm{IQR})$$

と定めて，この上下限(Fence)の間に含まれない変量(観測値)は，○や☆で表すことも一般的に行われています。



▶0.6.1 偏差 (deviation)

n 個の変量をもつデータ x の各値が，x_1，x_2，x_3，…，x_n であるとします。それらの相加平均(算術平均)が $\bar{x}$ であるとき，各変量(変数)と $\bar{x}$ の差の値，

$$x_1-\bar{x},\ x_2-\bar{x},\ x_3-\bar{x},\ \cdots,\ x_n-\bar{x}$$

を，平均値からの**偏差**とよび，各変量の平均値からのバラツキの数値を示します。

▶0.6.2 分散 (variance)

いかなるデータであっても，その偏差をもれなく足し合わせると '0' となります。

$$(x_1-\bar{x})+(x_2-\bar{x})+(x_3-\bar{x})+\cdots+(x_n-\bar{x})=0$$

したがって，各データに固有の要素(各値)のバラツキが評価できません。そこで，バラツキを定量化するために一般的に用いられる方法は，偏差の2乗を求め，その和を考えるという手法です。実際，

$$(x_1-\bar{x})^2,\ (x_2-\bar{x})^2,\ (x_3-\bar{x})^2,\ \cdots,\ (x_n-\bar{x})^2$$

は何れも0以上の値となり，バラツキの大きいデータの方が，この和が大きいという対応が存在します。

これらの総和をデータの個数 n で割ったもの(相加平均)

$$s^2=\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n}$$

を，**分散**といい，s^2 や V で表します。$(s\geqq 0)$

また，s すなわち，

$$s=\sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n}}$$

を，**標準偏差**(SD：standard deviation)といいます。厳密には，標本標準偏差とか試料標準偏差とよばれます。

また，変量xの平均値$\bar{x}$と，x^2の平均値$\overline{x^2}$を用いて，分散s^2は，次のように求めることができます。

$$s^2=\overline{x^2}-(\bar{x})^2$$

▶0.6.3　標準得点・標準化得点・z得点（standard score，z-score）

平均値も分布も異なる複数のデータを，同じ尺度で評価できたら便利です。そこで，データ内の変量x_k（$k=1, 2, \cdots, n$）が平均$\bar{x}$からどの位離れているかを定量するのに，偏差を標準偏差sで割った値である**標準得点**z_kを考えます。

$$z_k=\frac{x_k-\bar{x}}{s}$$

標準得点化されたデータの平均値は必ず'0'であり，標準偏差が'1'となります。

さらに，**偏差値**(t-score，deviation value) t_kとは，標準得点をA倍し，それにBを加えたものです。

$$t_k=\frac{x_k-\bar{x}}{s}\times A+B$$

▶0.6.4　変動係数（CV：coefficient of variation）

ある地域で競合する大型電気店において，LED電球(同一型番)とクリプトン電球(同一型番)の販売価格を調べたところ，LED電球の売価平均値は2,800円で標準偏差は120円だったのに対し，クリプトン電球の売価平均値は320円で，標準偏差は15円でした。しかしながら，標準偏差が大きいLED電球の方がクリプトン電球よりも価格の散らばり方が大きいという判断を即座にするのには問題がありそうです。そもそも高価格のものの方が価格のバラつきが大きいのは，地域内の洋菓子店におけるショートケーキとシュークリームの価格とか，スーパーにおけるメロンとバナナの値段などを考えてみれば自明とも思えることです。

そこで，平均値の大きく異なる複数のデータについて，その散らばり方の比較を図るために，標準偏差を平均値で割った**変動係数**という指標があります。

$$CV=\frac{s}{\bar{x}}$$

SAPIX YOZEMI GROUP