記述統計学のキーワード

記述統計学の“記述”とは，データの特徴を“記述する”ことを意味している。この場合，データが考察の対象とされる母集団全体であっても，母集団の中のごく限られた標本（サンプル）であっても構わない。単にそのデータの特徴を把握することに重点がおかれる。それに対して推測統計では，母集団から標本抽出された被験者のデータを用いて母集団全体の特徴を推測することが重視される。たとえば，被験者を選ぶ方法によっては，得られたデータが実際の母集団の特徴を全く反映していないこともある。このような母集団とのズレがどの程度のもので，それが実験誤差として許容範囲内であるのかどうかをもっともらしく判断するために考案されたのが，推測統計学であるといえるであろう。一般的な統計的分析では，記述統計学によって実験で得られたデータの特徴を把握し，推測統計学によってそこで示された特徴の頑健性を検証するという手続きをとる。

なお，統計的処理に慣れてくると，グラフを作成するなどデータの記述的特徴の把握を無視していきなり差異検定などを行う人もでてくるが，これは推論を誤らせる可能性が高く避けるべきである。たとえば，あるテストAの得点と別のテストB，テストCの得点とが似たような数値である場合，同じような得点のテスト（e.g., 同じような性格特性を測定するテスト）として一括りに考えて良いかどうかは問題がある。むしろ，テストAに対する相関の仕方から判断した場合には，テストBとテストDとは平均点こそ大きく異なるが，似たようなテストであると考えることができる（図1と図3）。逆に，テストBとテストCとは平均値こそよく似ているが，同種のテストと考えるのには無理があるであろう（図1と図2）。実際に今回サンプルとして提示したデータは以下のような手順で作成された。テストBは40～100の間のランダムな数値で，この値と更に別のランダムな数値（50～80）を足して平均した値がテストAとなっている。それに対して，テストCは，テストBの値の40～60の範囲の値から一様に20引き，70～90の範囲の値に20を足して，極端に上下に分布が別れるように作成した数字列となっている。また，テストDはテストBの値を単純に2で割って四捨五入した値となっている。それぞれのデータについて記述統計量を求めて比較検討されたい。

最低限知っておきたいキーワード

大量にあるデータがどのような性質をもっているかを把握するために，少数の代表値を用いることがある。たとえば，平均値（mean）や，中央値（median），最頻値（mode）などである。しかしながら，これらの代表値だけではデータの特徴を記述するのに不十分である。たとえば，ある条件とある条件との平均値が大きく異なっているように見えた場合にも，一方の条件にハズレ値があってそのため平均が異なっていたならば，条件間で差があるとは言い難い。このような不具合を解消するために，データの分布の特徴を記述する方法として中心的データである平均値からの散らばり具合を把握する必要があり，その代表的な値として分散（variance または s^2）ならびに標準偏差（Standard Deviation, STD または s，S(x)）が用いられる。平均値と分散は以下の式によって定義される。x_i とは全体でN人いる被験者のi人目の被験者のデータの意味である。

平均値… $Mean = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$　　分散… $Variance = s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$

なお，散布度を表す値として，平均との差の絶対値を用いることも良さそうに見えるかも知れないが，絶対値を用いた計算は煩雑になりやすく式の展開がしにくいという欠点があるため，上記の平均との差の二乗和が採用されている．また，全体を被験者の人数であるNで割るのは，標本の大きさに依存しない指標である必要があるからである．分散はズレが二乗されているので，その平方根をとったものである標準偏差が散らばり具合を代表する値として用いられることもある．

1．**尺度の種類**…大まかに平均値を求めて意味のある尺度（間隔尺度，比率尺度）とそうでない尺度（順序尺度，名義尺度）とに分類できる。後者の尺度に基づくデータは，質的データ，カテゴリカルなデータ，属性データと呼ばれ，χ 二乗検定に見られるノンパラメトリカル（non-parametoric）な検定の対象とされることが多い。一般的な分類としては以下の表の通りになる。

尺度水準	例	説明
名義尺度	性別，所属クラブ，専攻	たとえ０，１といった具合に数字が振られていたとしても，加算したり，平均値を求めることが無意味なデータ。
順序尺度	好きな順位や美人コンテストの順位など	数値の大小関係は順序関係を反映しているが，その差が等間隔であるかどうかの保証がない。平均値よりも中央値を求める方が本来は相応しい。
間隔尺度	知能偏差，暦年など	数値の差の間隔が等しい。ただし，原点および単位は任意の変換操作が可能。
比例尺度	長さ，質量，反応時間	原点が存在し，差に加えて比が意味を持つ。

２．**主要代表値**…平均値，中央値，モード（最頻値），調和平均

平均値の特徴

a. 平均値からの全ての得点の差の集計が 0 になる，もっとも均衡のとれた点（重心）。
b. 比較的左右対称な分布の測度としてふさわしい。中央値や最頻値よりも極端な値に影響されやすく，歪んだ分布に適応するのは好ましくない。
c. 数学的に操作しやすいが，質的データに適応するのは好ましくない。

中央値の特徴

a. 分布の特定の値に支配されないため，安定性の高い測度である。
b. 中央値の上下の値に敏感ではないから，著しく歪んだ分布にも使用できる。
c. 数理的な操作が面倒である。

最頻値

a. もっとも頻繁に生起する値という意味でもっとも典型的な値である。
b. 計算も簡便で理解しやすい。
c. 数理的な操作に向かない。
d. 離散的データよりも質的データの分析に適している。

３．**主要散布度**…分散，標準偏差，四分位数，歪度・尖度（この中でも分散と標準偏差が特に重要）

とても重要な分布…正規分布（normal distribution）

正規分布をするデータでは，平均値，中央値，最頻値が一致し歪度は 0 となる。逆に言えば，それ以外の分布では一致しないことの方が多いといえる。正規分布の重要な特徴として以下の 4 点が挙げられる。

１．平均と標準偏差が決まれば分布の形も決まる。
２．平均や標準偏差が変わっても，分布のタイプ（正規性）は変わらない（歪度＝0，尖度＝3 が維持される）。
３．平均と標準偏差が分かれば，任意の得点間に入る相対頻度が計算できる（→平均±1 標準偏差の区間に入る頻度は 68%）。
４．得点に対して線形変換を行っても分布の正規性が保たれる（→ z 変換，標準化）。

得点の標準化（正規化）$\rightarrow z_i = \dfrac{x_i - \bar{x}}{S(x)}$，

$\rightarrow z_i$ の値の平均が 0 で標準偏差が 1 の正規分布になる。最も重要なことは，返還後の値が特定の単位系に依存しない値に変換されることである。たとえば，㎝で計ったものを㎝で割ることによって㎝の単位が相殺される。同様に㎏で測られた値も，標準化すれば㎏に依存しない値に変換できる。その結果，両者の値を同じ基準で比較できるようになる（そのような比較に意味があるかは別として）。

偏差値得点の計算＝$10\dfrac{x_i - \bar{x}}{S(x)} + 50$

→平均が 50 で標準偏差が 10 の正規分布になる。たとえば，偏差値 60 とは上位約 15%の区間に入ることを示す。

	テストA	テストB	テストC	テストD
	64	48	28	24
	81	88	88	44
	60	57	37	29
	58	45	25	23
	80	97	97	49
	67	60	80	30
	70	80	80	40
	82	90	90	45
	80	88	88	44
	72	70	90	35
	73	66	86	33
	56	52	32	26
	73	83	83	42
	69	77	97	39
	87	100	100	50
	74	70	90	35
	81	94	94	47
	77	89	89	45
	60	54	34	27
	72	79	99	40
	75	79	99	40
	79	86	86	43
	69	60	30	30
	70	71	91	36
	62	43	23	22
	55	50	50	25
	71	62	82	31
	75	73	93	37
	73	81	81	41
	79	95	95	48
	63	46	26	23
	73	72	92	36
	58	52	32	26
	74	70	90	35
	79	91	91	46
	68	72	90	36
	81	85	85	43
	82	85	85	43
	61	55	35	28
	81	94	94	47
	61	58	38	29
	79	91	91	46
	60	41	21	21
	64	63	83	32
	84	89	89	45
	72	67	87	34
	80	96	96	48
	81	90	90	45
	74	85	85	43
	57	49	29	25
平均	71.52	72.76	73.32	36.62
標準偏差	8.52	16.84	26.78	8.44

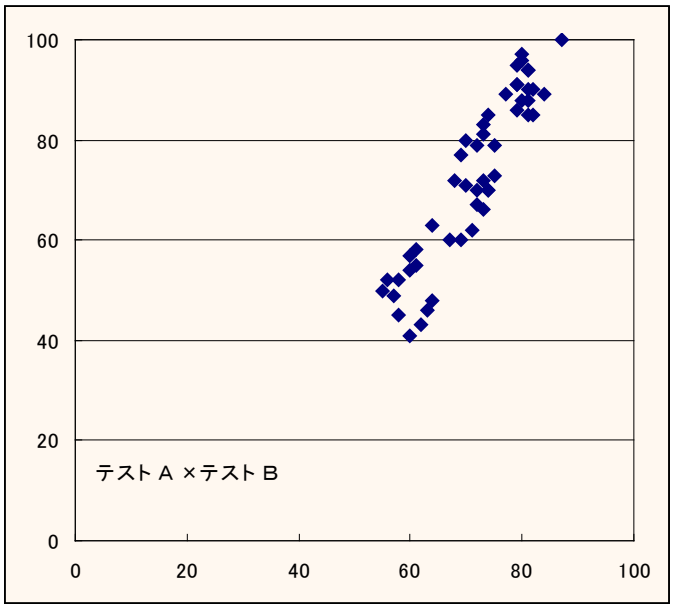
100
80
60
40
20
0
テスト A ×テスト B
0
20
40
60
80
100

図1

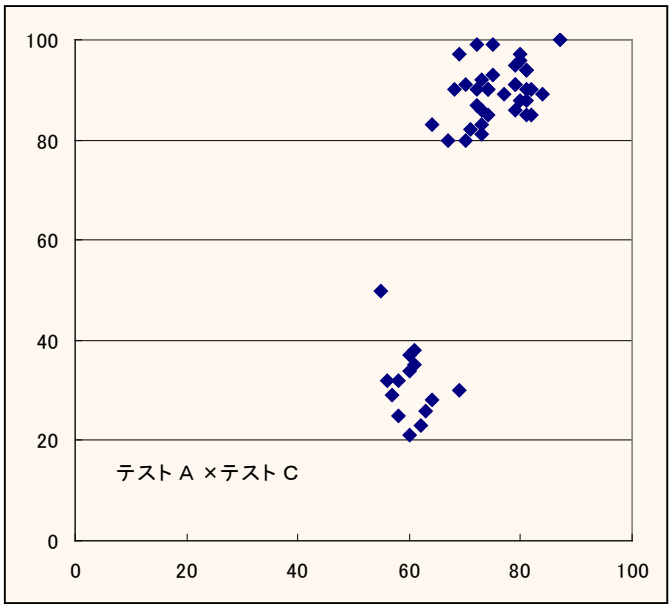
100
80
60
40
20
0
テスト A ×テスト C
0
20
40
60
80
100

図2

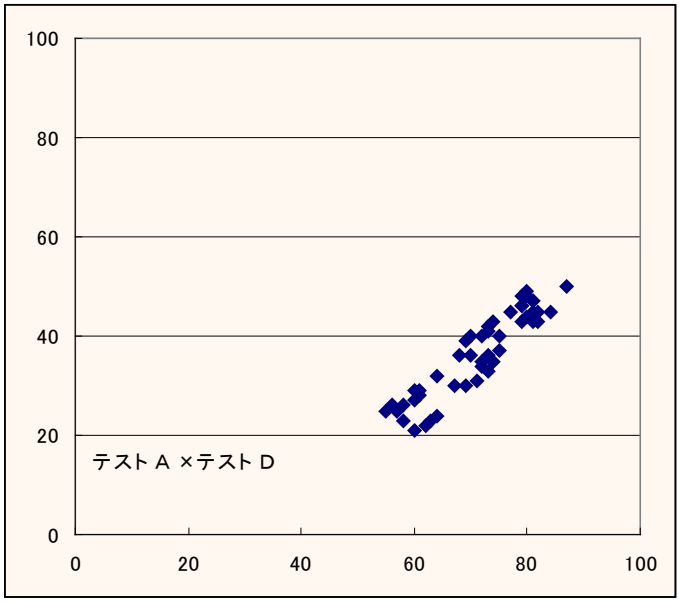
100
80
60
40
20
0
テスト A ×テスト D
0
20
40
60
80
100

図3