データサイエンスのための統計学

1 次元のデータ

by 尚晋　(名古屋大学経済学研究科助教)
on 2025 年 4 月 22 日

目次

* 度数分布とヒストグラム
* 記述統計量: 位置の尺度
* 記述統計量: 散らばりの尺度

目次

* 度数分布とヒストグラム
* 記述統計量: 位置の尺度
* 記述統計量: 散らばりの尺度

» 記述統計学に関して

記述統計学とは　集団としての特徴を記述するために，観測対象となった各個体について観測し，得られたデータを＿＿＿＿＿＿ する方法である．

観測とは　広く調査や実験のこと．

データとは　各「個体」(人，もの) の＿＿＿＿＿ をまとめたもの．

» 記述統計学に関して

記述統計学とは　集団としての特徴を記述するために，観測対象となった各個体について観測し，得られたデータを整理・要約 する方法である．

観測とは　広く調査や実験のこと．

データとは　各「個体」(人，もの) の観測値 をまとめたもの．

» 度数分布

- 調査や実験によって観測値が得られたとき, 最初に度数分布表を作ることから始める.
- 計算するよりも, _____ にする方が全体の分布の状況が明らかになるからである.
- 例えば、ある大学における統計学の試験の受験者数 373 人の成績を度数分布表にしたのが, 表 2.1 である.

» 度数分布

- 調査や実験によって観測値が得られたとき, 最初に度数分布表を作ることから始める.
- 計算するよりも, 表や図 にする方が全体の分布の状況が明らかになるからである.
- 例えば、ある大学における統計学の試験の受験者数 373 人の成績を度数分布表にしたのが, 表 2.1 である.

» 度数分布

度数分布表

表 2.1 試験得点の度数分布表 (某大学の統計学)

階	級	階級値	度数	相対度数	累積度数	累積相対度数
0≤	<10	5	12	0.032	12	0.032
10"	20"	15	10	0.027	22	0.059
20"	30"	25	19	0.051	41	0.110
30"	40"	35	42	0.113	83	0.223
40"	50"	45	72	0.193	155	0.416
50"	60"	55	82	0.220	237	0.635
60"	70"	65	54	0.145	291	0.780
70"	80"	75	38	0.102	329	0.882
80"	90"	85	25	0.067	354	0.949
90"	≤100	95	19	0.051	373	1.000
合	計		373	1.000		

» 度数分布

度数分布に関する定義

* 度数分布表は, 観測値のとりうる値をいくつかの＿＿に分け, それぞれの階級で観測値がいくつあるか＿＿を数えて, 表にしたものである.
* ＿＿＿＿とは階級を代表する値のことであって, 階級の上限値と下限値の中間値を階級値とするのが普通である.
* (度数)/(観測値の総数) を＿＿＿＿＿という.
* 累積度数, 累積相対度数とは, 度数を下の階級から順に積み上げたときの度数, 相対度数の＿＿＿＿ である.

» 度数分布

度数分布に関する定義

* 度数分布表は, 観測値のとりうる値をいくつかの階級に分け, それぞれの階級で観測値がいくつあるか度数を数えて, 表にしたものである.
* 階級値とは階級を代表する値のことであって, 階級の上限値と下限値の中間値を階級値とするのが普通である.
* (度数)/(観測値の総数) を相対度数という.
* 累積度数, 累積相対度数とは, 度数を下の階級から順に積み上げたときの度数, 相対度数の累積和 である.

» ヒストグラム

- 横軸に観測値のとりうる値, 各階級に対して階級幅を横幅とし, 柱の面積が各階級の (相対) 度数と比例するように高さを定める, このようなグラフを________________という.

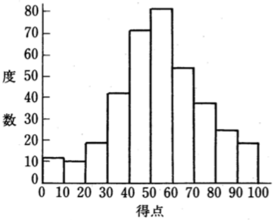

図 2.1　試験得点のヒストグラム
度数分布がこのように整った形となることはむしろめずらしい. 階級数, 階級幅などをうまくとらねばならない.

図 **2.1** 試験得点のヒストグラム

- 試験の成績の分布は, 図に示されているように, 中央に一つ峰がある山型分布である.

» ヒストグラム

* 横軸に観測値のとりうる値, 各階級に対して階級幅を横幅とし, 柱の面積が各階級の (相対) 度数と比例するように高さを定める, このようなグラフをヒストグラム (柱状グラフ)という.

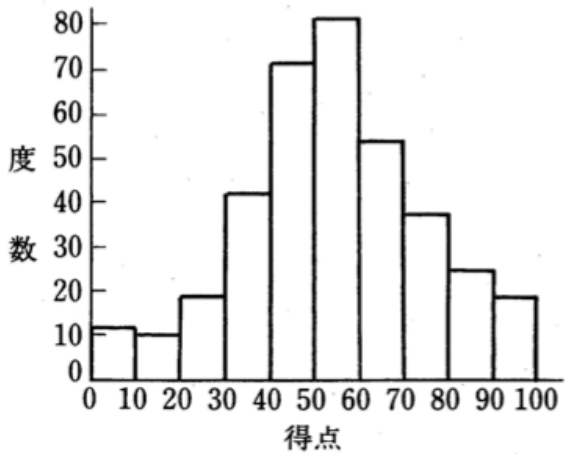

図 2.1　試験得点のヒストグラム
度数分布がこのように整った形となることはむしろめずらしい. 階級数, 階級幅などをうまくとらねばならない.

図 **2.1** 試験得点のヒストグラム

* 試験の成績の分布は, 図に示されているように, 中央に一つ峰がある山型分布である.

» ヒストグラム

右に歪んだ分布

* しかし, 左右対称の山型分布にならないものも多くある. そのうち峰が中央から左側に寄っていて, 右側に長く裾を引く分布のことを, (感覚とは逆になるが)_____ という. 例:図 2.2

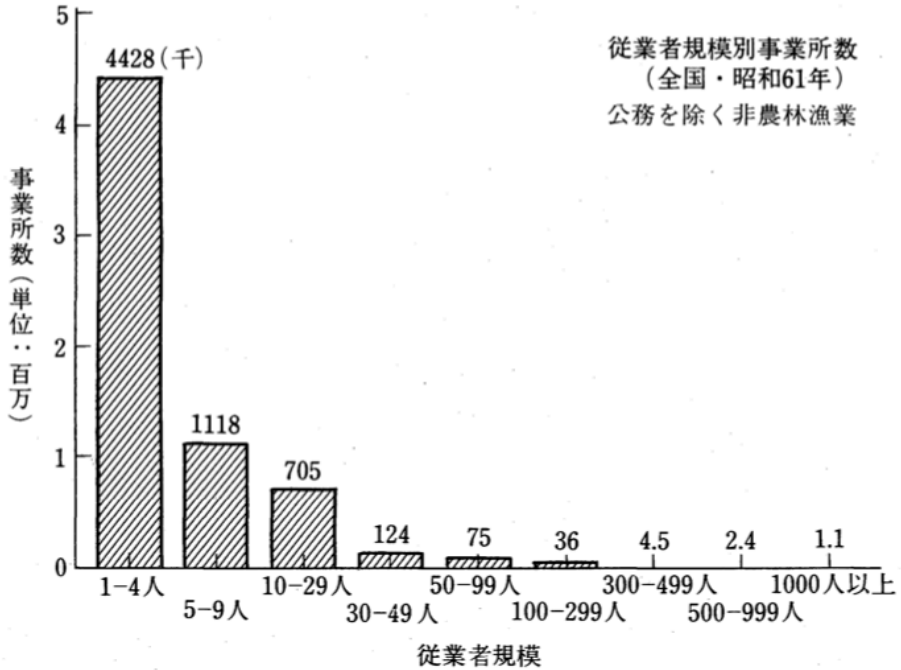

図 2.2　従業者規模別事業所件数(全国・1986 年)

峰が左端に寄り，右に長く尾をひいた分布(右に歪んだ分布)の例である．この図のように，階級幅が各階級で著しく異なる場合には，柱を分離して描く．

(出典：総務庁統計局「事業所統計調査報告」)

» ヒストグラム

右に歪んだ分布

* しかし, 左右対称の山型分布にならないものも多くある. そのうち峰が中央から左側に寄っていて, 右側に長く裾を引く分布のことを, (感覚とは逆になるが)右に歪んだ分布 という. 例: 図 2.2

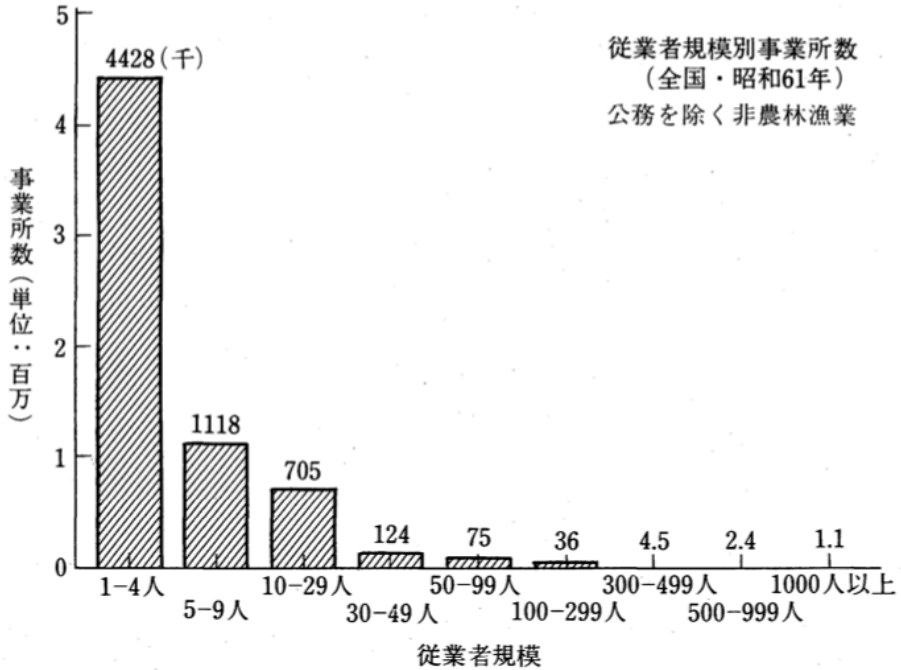

図 2.2　従業者規模別事業所件数(全国・1986 年)

峰が左端に寄り，右に長く尾をひいた分布(右に歪んだ分布)の例である．この図のように，階級幅が各階級で著しく異なる場合には，柱を分離して描く．

(出典：総務庁統計局「事業所統計調査報告」)

» ヒストグラム

階級数の問題と階級幅の問題

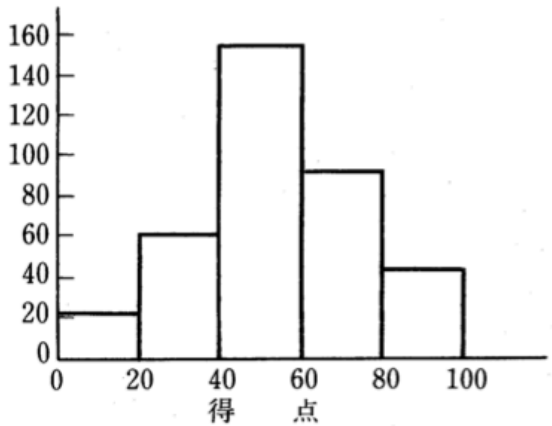

図 2.6 試験得点のヒストグラム
(階級数が少ない場合)

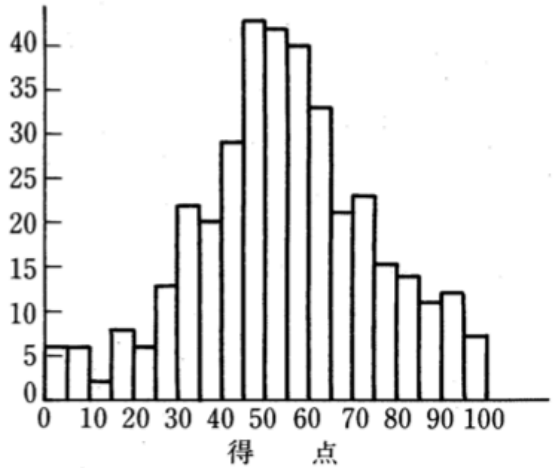

図 2.7 試験得点のヒストグラム
(階級数が多い場合：その 1)

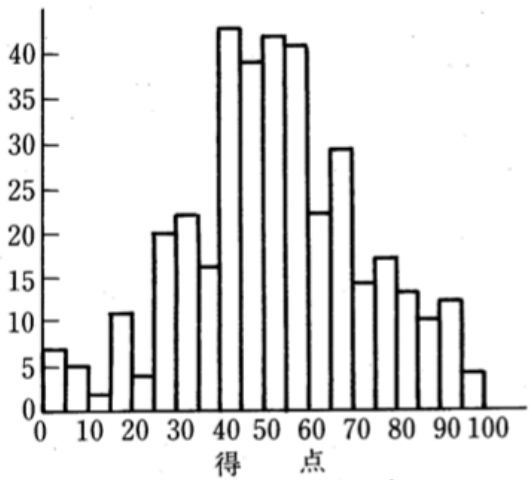

図 2.8 試験得点のヒストグラム
(階級数が多い場合：その 2)

図 2.1 を含むこれら四つは，全て同一のデータからである．度数分布の多様性とともに，'難しさ' もわかるであろう．

» ヒストグラム

階級数の問題と階級幅の問題

- 度数分布表やヒストグラムを作成するときに注意すべき点は，階級数の問題と階級幅の問題である．例：図 2.6, 図 2.7, 図 2.8
- きざみが＿＿＿＿＿ なら，真の分布を見出しえない．階級幅を＿＿＿＿ とり，階級数を増やすと，分布が階級のとり方に敏感になる．
- 階級数に関してはスタージェスの公式が参考になる．観測値の数を n とし，階級数 k とした時の式は：
  $k \fallingdotseq 1 + log_2 n = 1 + (log_{10} n)/(log_{10} 2)$
- 階級幅に厳密な決まりはないが、通常は等しい幅が望ましい。ただし、分布の端で度数が極端に少ない場合は、階級幅を広げることがある。

» ヒストグラム

階級数の問題と階級幅の問題

- 度数分布表やヒストグラムを作成するときに注意すべき点は，階級数の問題と階級幅の問題である．例：図 2.6, 図 2.7, 図 2.8
- きざみが粗すぎる なら，真の分布を見出しえない．階級幅を小さく とり，階級数を増やすと，分布が階級のとり方に敏感になる．
- 階級数に関してはスタージェスの公式が参考になる．観測値の数を n とし，階級数 k とした時の式は：
  $k \fallingdotseq 1 + log_2 n = 1 + (log_{10} n)/(log_{10} 2)$
- 階級幅に厳密な決まりはないが、通常は等しい幅が望ましい。ただし、分布の端で度数が極端に少ない場合は、階級幅を広げることがある。

» 累積度数のグラフ

* 度数分布表からは累積度数や累積相対度数をもとにしたグラフをつくることもできる. 累積度数グラフの例：図 2.11

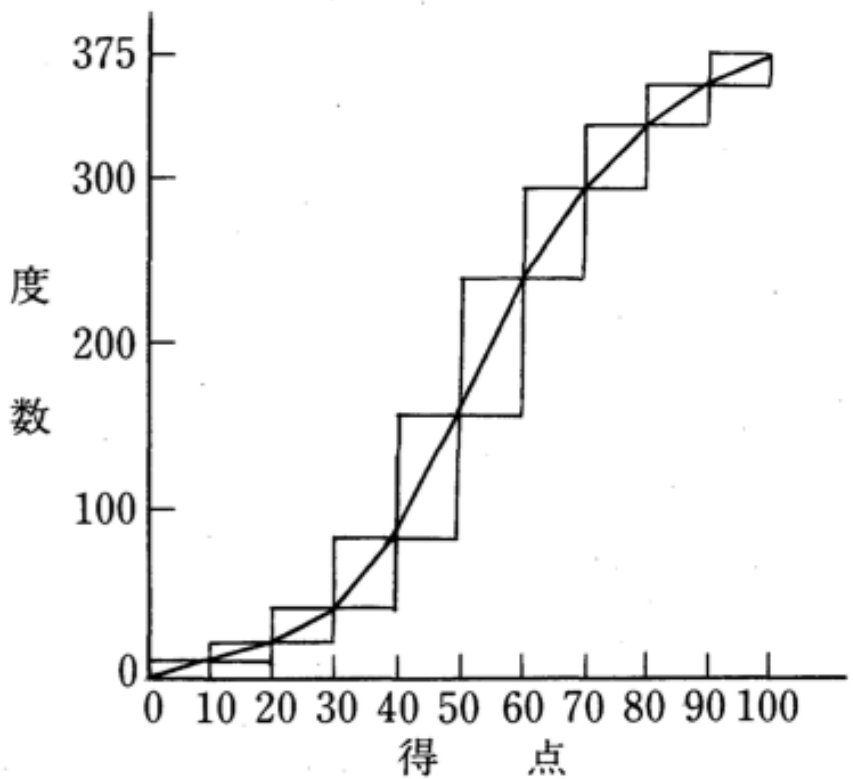

図 2.11　ヒストグラムと累積度数グラフ

» ローレンツ曲線

表 2.3　従業者規模別事業所数および従業者数
(全国・1986 年・公務を除く非農林漁業)

従業者規模	事業所数 (千件)	累積相対度数	従業者数 (千人)	累積相対度数
1- 4人	4,428	0.682	9,486	0.194
5- 9人	1,118	0.854	7,214	0.341
10- 29人	705	0.963	11,134	0.568
30- 49人	124	0.982	4,648	0.663
50- 99人	75	0.993	5,103	0.767
100-299人	36	0.999	5,734	0.884
300-499人	4.5	0.999	1,706	0.919
500-999人	2.4	1.000	1,651	0.953
1000人以上	1.1	1.000	2,320	1.000
合　計	6,494		48,995	

(出典：総務庁統計局「事業所統計調査報告」)

表 2.3: 事業所数と従業者数

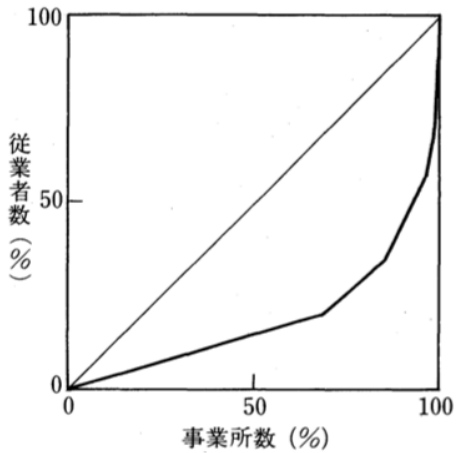

図 2.13　事業所規模のローレンツ曲線
(全国・昭和 61 年)

図 2.13: 事業所規模のローレンツ曲線

» ローレンツ曲線

* 累積相対度数のグラフでは、異なる 2 つのデータ（例：表 2.3 の事業所数と従業者数）を組み合わせて表すことができる。
* 例えば、表 2.3 のデータに基づいて，横軸に事業所数、縦軸に従業者数の累積相対度数を取り、各点を線で結ぶと図 2.13 のようなグラフが作成できる。
* これは＿＿＿＿＿＿＿＿であり、事業所数の最初の何%に従業者数の何%が含まれるかを示すグラフである。
* 全ての事業所に同じ従業者数がいれば、線は対角線になる。対角線からのずれが大きいほど、規模の＿＿＿＿ が大きい。ローレンツ曲線は、所得や資産の不平等を示すのにも使われる。

» ローレンツ曲線

* 累積相対度数のグラフでは、異なる 2 つのデータ（例：表 2.3 の事業所数と従業者数）を組み合わせて表すことができる。
* 例えば、表 2.3 のデータに基づいて，横軸に事業所数、縦軸に従業者数の累積相対度数を取り、各点を線で結ぶと図 2.13 のようなグラフが作成できる。
* これはローレンツ曲線であり、事業所数の最初の何%に従業者数の何%が含まれるかを示すグラフである。
* 全ての事業所に同じ従業者数がいれば、線は対角線になる。対角線からのずれが大きいほど、規模の不均等 が大きい。ローレンツ曲線は、所得や資産の不平等を示すのにも使われる。

目次

* 度数分布とヒストグラム
* 記述統計量: 位置の尺度
* 記述統計量: 散らばりの尺度

» 位置の尺度

_____ とは (観測値の総和) / (観測値の総数)、即ち観測値の総和を観測値の総数で割ったもの.

_____ とは 観測値を小さい方から順に並べたときの中央の値.

分位点とは 観測値を小さいものの順に並びかえたとき, 小さい方から $100p\%(0 \leqq p \leqq 1)$ の所にある値を _____ という. よく用いられる分位点には四分位点がある: 第 1 四分位点 Q_1 は 25% 分位点, 第 2 四分位点 Q_2 は 50% 分位点 (メディアン), 第 3 四分位点 Q_3 は 75% 分位点

_____ とは その度数が最大である階級の階級値のこと.

» 位置の尺度

算術平均 とは (観測値の総和) / (観測値の総数)、即ち観測値の総和を観測値の総数で割ったもの.

中位数 (中央値或いはメディアン) とは 観測値を小さい方から順に並べたときの中央の値.

分位点とは 観測値を小さいものの順に並びかえたとき, 小さい方から $100p\%(0 \leqq p \leqq 1)$ の所にある値を$100p$ パーセンタイルまたは分位点 という. よく用いられる分位点には四分位点がある: 第 1 四分位点 Q_1 は 25% 分位点, 第 2 四分位点 Q_2 は 50% 分位点 (メディアン), 第 3 四分位点 Q_3 は 75% 分位点

モード (最頻値) とは その度数が最大である階級の階級値のこと.

目次

* 度数分布とヒストグラム
* 記述統計量: 位置の尺度
* 記述統計量: 散らばりの尺度

» 散らばりの尺度

導入問題

大きさが同じ $n = 10$ の以下 3 つのデータ A,B,C があったとする. これらの平均, メディアン, モードはいずれも 5 である.

A:	0	3	3	5	5	5	5	7	7	10
B:	0	1	2	3	5	5	7	8	9	10
C:	3	4	4	5	5	5	5	6	6	7

- この $A, B, C3$ つの分布は, なめらかな分布を想定すると図 2.17 のような関係になっている.
- 前節で説明した代表値は分布の位置を示す指標であって, 平均, メディアン, モードはいずれも同じ.
- この例のような $A, B, C3$ つの分布を区別するには, 分布の形状を示す他の指標が必要となることがわかる.

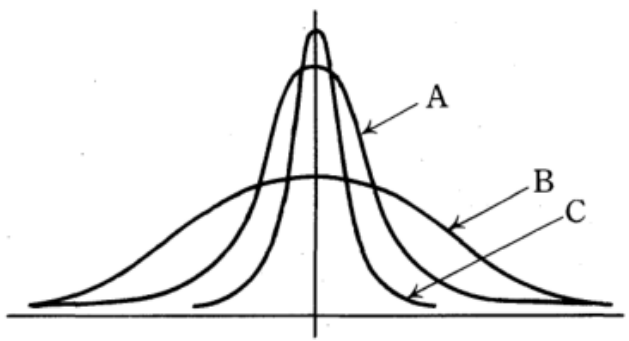

図 **2.17:** 同一平均で異なる分布図の **A,B,C**

» 散らばりの尺度

分布の形状を示す指標は多くある. 散らばりの尺度と呼ばれるものはよく使われる。位置の尺度と散らばりの尺度の二つを用いれば, 分布のおおまかな形状を記述することができる.

_____ とは もっとも単純, 分布の存在する範囲を示すもの。最大値と最小値の差と計算する。かなり粗いものである. あまり使わない.

_____ とは データの第 3 四分位点 Q_3 と第 1 四分位点 Q_1 の隔たりの半分として定義され, 真中の半分のデータが散らばっている範囲の平均を表す. 四分位偏差が大きいほど散らばった分布となる.

* レンジも四分位偏差も, たかだか 2 個ないし 4 個の観測値を用いるだけ.

» 散らばりの尺度

分布の形状を示す指標は多くある. 散らばりの尺度と呼ばれるものはよく使われる。位置の尺度と散らばりの尺度の二つを用いれば, 分布のおおまかな形状を記述することができる.

レンジ (範囲) とは　もっとも単純, 分布の存在する範囲を示すもの。最大値と最小値の差と計算する。かなり粗いものである. あまり使わない.

四分位偏差 とは　データの第 3 四分位点 Q_3 と第 1 四分位点 Q_1 の隔たりの半分として定義され, 真中の半分のデータが散らばっている範囲の平均を表す. 四分位偏差が大きいほど散らばった分布となる.

* レンジも四分位偏差も, たかだか 2 個ないし 4 個の観測値を用いるだけ.

» 散らばりの尺度

平均偏差, 標準偏差はすべての観測値のもつ情報を利用した散らばりの尺度, いずれも各観測値 x_i と平均 $\bar{x}$ との隔たり (＿＿＿ という) をもとに計算される.

＿＿＿＿ とは 各観測値が平均からどれだけ離れているかについての平均を求めたもの。平均偏差の計算式:
$d = \frac{1}{n}\{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \ldots + |x_n - \bar{x}|\}$

＿＿＿ とは 平均からの偏差の 2 乗の平均. 分散は S^2 という記号で表され, 分散の計算式:
$S^2 = \frac{1}{n}\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2\} = \frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})^2$

＿＿＿＿ とは 分散の平方根.

» 散らばりの尺度

平均偏差, 標準偏差はすべての観測値のもつ情報を利用した散らばりの尺度, いずれも各観測値 x_i と平均 $\bar{x}$ との隔たり（偏差 という）をもとに計算される.

平均偏差 とは　各観測値が平均からどれだけ離れているかについての平均を求めたもの。平均偏差の計算式:

$$d = \frac{1}{n}\{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \ldots + |x_n - \bar{x}|\}$$

分散 とは　平均からの偏差の 2 乗の平均. 分散は S^2 という記号で表され, 分散の計算式:

$$S^2 = \frac{1}{n}\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2\} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

標準偏差 とは　分散の平方根.

» 変動係数

導入問題

導入問題

例えば, 1965 年には 1 人あたり県民所得の平均は $\bar{x} = 26.6$ 万円, 標準偏差は $S = 7.5$ 万円であったのが, 1975 年には平均 $\bar{x} = 117.5$ 万円, 標準偏差 $S = 23.8$ 万円になっている. 地域間の所得格差は大きくなっているのであろうか.

- 分布の中心の位置が, 著しく異なるような場合には, 分散 (標準偏差) をもって分布の散らばり具合を比較することはできない.
- この場合は＿＿＿＿＿＿を用いて比較する.

変動係数とは (標準偏差) / (平均). 変動係数はよく C.V. という記号で表され、計算式は:

$$C.V. = \frac{S_x}{\bar{x}}$$

» 変動係数

導入問題

導入問題

例えば, 1965 年には 1 人あたり県民所得の平均は $\bar{x} = 26.6$ 万円, 標準偏差は $S = 7.5$ 万円であったのが, 1975 年には平均 $\bar{x} = 117.5$ 万円, 標準偏差 $S = 23.8$ 万円になっている. 地域間の所得格差は大きくなっているのであろうか.

- 分布の中心の位置が, 著しく異なるような場合には, 分散 (標準偏差) をもって分布の散らばり具合を比較することはできない.
- この場合は変動係数を用いて比較する.

変動係数とは (標準偏差) / (平均). 変動係数はよく C.V. という記号で表され、計算式は:

$$C.V. = \frac{S_x}{\bar{x}}$$

» 変動係数

* 単純に標準偏差を比較すれば約 3 倍になっており, 大きくなっている. しかしその間に平均も約 4.5 倍に増えている. この例のように直接の比較が困難な場合に, 平均えを考慮した上で散らばり具合を相対的に比較するのに便利な指標である.

  この例では 1965 年の変動係数は _____,
  1975 年は _____ であり,

  結果の解釈:

  _____.

» 変動係数

* 単純に標準偏差を比較すれば約 3 倍になっており, 大きくなっている. しかしその間に平均も約 4.5 倍に増えている. この例のように直接の比較が困難な場合に, 平均えを考慮した上で散らばり具合を相対的に比較するのに便利な指標である.

  この例では 1965 年の変動係数は $7.5/26.6 = 0.28(28\%)$,
  1975 年は$23.8/117.5 = 0.20(20\%)$ であり,

  結果の解釈:
  安定している中にも相対的な地域間所得格差はむしろ小さくなっている.

» 標準得点 (標準化)

_____ とは 平均を差し引き, 標準偏差で割って, 位置, 尺度の調整をした結果, 平均は $\bar{z} = 0$, 標準偏差は $S_z = 1$ に揃ったことになる. この z をデータ x の標準化とか, 標準得点という. 計算式は:

$$z_i = \frac{x_i - \bar{x}}{S_x}$$

_____ とは 試験の得点 (スコア) を平均が $\bar{z} = 50$ 点, 標準偏差が $S_z = 10$ 点となるように変換したもの. 偏差値得点 T_i は:

$$T_i = 10z_i + 50$$

で計算する. z_i, T_i はそれぞれ Z 得点, T 得点と呼ばれる.

例: 標準得点と偏差値

例えば, あるのテストを受けた 200 人の得点は、平均点が 58 点で標準偏差が 8 であった. このテストで 50 点だった人の標準得点と偏差値得点を求めよ. 答えは_____.

» 標準得点 (標準化)

標準得点 (標準化) とは　平均を差し引き, 標準偏差で割って, 位置, 尺度の調整をした結果, 平均は $\bar{z} = 0$, 標準偏差は $S_z = 1$ に揃ったことになる. この z をデータ x の標準化とか, 標準得点という. 計算式は:

$$z_i = \frac{x_i - \bar{x}}{S_x}$$

偏差値得点 とは　試験の得点 (スコア) を平均が $\bar{z} = 50$ 点, 標準偏差が $S_z = 10$ 点となるように変換したもの. 偏差値得点 T_i は:

$$T_i = 10z_i + 50$$

で計算する. z_i, T_i はそれぞれ Z 得点, T 得点と呼ばれる.

例: 標準得点と偏差値

例えば, あるのテストを受けた 200 人の得点は、平均点が 58 点で標準偏差が 8 であった. このテストで 50 点だった人の標準得点と偏差値得点を求めよ. 答えは標準得点は -1 で, 偏差値得点は 40.