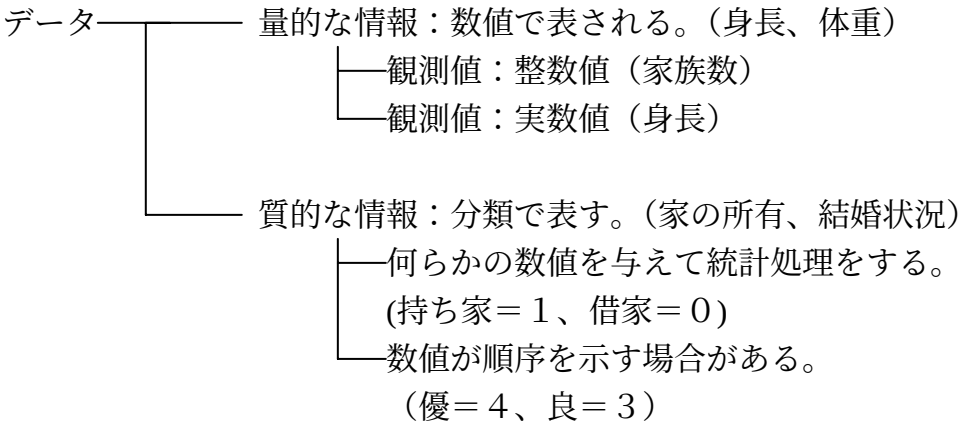
統計科学 (Statistics：講義 2)

１．データの整理

データ

調査や実験からえられた情報。



1.1 データの代表値

母集団：調査対象の全体（ある大学における在学生の月々の所得）

標本（データ）：一部の在学生

母集団全体が観測された場合の観測値の集合を全標本(センサス census)という。

変数 X：在学生の月々の所得

計測値：$x_1, x_2, ..., x_n$.

計測値の集まり：　$S = \{x_1, x_2, ..., x_n\}$.

所得を調べた学生の総数：観測個数 n

統計学では標本を遣って母集団全体を理解しようとする。このための思考方法を**統計的推測**とよぶ。

データ全体の中心を表す尺度

1．標本平均値 (mean)

観測値の総和を観測個数で割った値。

$$mean: \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Note: Population mean: μ or m.

２．中央値 (median)　あるいは中位数

観測値を小から大（もしくは大から小）まで順番に並べたときの「真ん中」の値である。

観測値が奇数：真ん中の値。

観測値が偶数：真ん中二つの平均。

３．最頻値 (mode)

最も頻繁に現れた値。

例：　$S = \{3,7,2,4,3,8,10,5,3,5\}$

mean= 5　; median=　(4+5)/2=4.5　; and mode=　3 .

例：　$S = \{3,7,2,4,3,5,8,5,3,5\}$ では、次の値はなんですか。

mean=　　; median=　; and mode=　.

データの広がりを示す代表値

１．標本分散 (sample variance)

データの中心のまわりでの散らばりを示す。「平均値から測った距離の２乗」の平均である。

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \quad .$$

なぜ$n-1$で割るのかは、統計学理論上、unbiased estimator であるから。nが大きくなればなるほど、$n-1$でもnで割っても結果に大きな差は生じない。

$s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$ とした場合に、

$$s^2 = \left(\frac{1}{n}\right)\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \left(\frac{1}{n}\right)\left[\sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\bar{x}^2\right]$$

$$= \left(\frac{1}{n}\right)\left[\sum_{i=1}^{n}x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right] = \left(\frac{1}{n}\right)\left[\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right] = \left[\frac{\sum_{i=1}^{n}x_i^2}{n} - \bar{x}^2\right], where \sum_{i=1}^{n}x_i = n\bar{x}.$$

母集団全体（全標本）の分散 (population variance)

$\sigma^2 = \dfrac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$　もしくは　$\sigma^2 = \dfrac{\sum_{i=1}^{n}(x_i - m)^2}{N}$ ．

２．標本標準偏差 (sample standard deviation)

標本分散の平方根

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad .$$

母集団の標準偏差（分散の平方根）

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}} \quad .$$

標準偏差はデータの広がりを示す重要な統計量（またはパラメータ）である。

<u>チェビシェフの不等式</u>

１よりも大きなk について、標本平均を囲む区間

（標本平均値―k x標本標準偏差）～（標本平均値＋k x標本標準偏差）

に入らない観測値は、全体の$\dfrac{1}{k^2}$以下である。

例：$k = 2$ の場合は、$\dfrac{1}{2^2} = \dfrac{1}{4} = .25(= 25\%)$ が、入らない観測値の割合で、この区間に入る観測値は 75％となる。

記述統計量

標本平均値や分散および標準偏差などデータの性質を表す基本的なものをさす。しかし、これらは異常値に左右される。一方、中央値や最頻値はこの限りでない。

３．四分位点(quartile)（四分位階級）

観測値全体を小から大に並べたとき、

２５％目の値：２５％点、つまり第１四分位点

５０％目の値：中央値

７５％目の値：第３四分位点

四分位範囲＝第３四分位点－第１四分位点.

四分位分散係数＝四分位範囲を中位数の２倍で割った比

十分位点（十分位階級）

１０％ごとの値

1.2 度数分布表

母集団の性質を知るために、母集団から集められた標本から**中心を示す代表値**（平均値、中央値、最頻値）や**中心（平均値）からの広がり**（標本分散、標本標準偏差）を得ることが、良い方法である。しかし、観測個数nが多い（つまり観測値x_iが観測個数の数だけある）と、一目では標本の観測値から知ることは困難である。

例えば、$n=100$では、100 個の観測値を机の上に並べて見たとする。で、何か分かるでしょうか？おそらく、どの数字（観測値）が多いなと知ることまでは出来るでしょう。また、一番小さな値は？一番大きな値は？など興味をもって 1 時間ほど眺めているでしょう。

上記の方法も良いけど、なんとなく整理をしたくなるというのが、悲しい人間の性格なのでしょう。その観測値を整理する方法はいろいろあるでしょうが、ここでは**度数分布表**の作り方を学びます。

度数分布表 (Frequency Table)

観測値を大きさによっていくつかのグループに分けて、それぞれの観測値x_iをそれらのグループに割り振った表を、度数分布表と言います。

では、いくつのグループに割り振ったらよいのでしょうか？

答え：　グループの数は観測個数nによります。

Step 1:　このグループの数を**階級数**と呼びます。その階級数の数を決める一つの方法（スタージェス (Starjes)の公式、無理して覚えなくても、度数分布表の作り方を覚えるほうが大切！）は：

$$\text{階級数}=1+3.3x\log_{10}n \quad (\text{対数}\log_{10}\text{は常用対数})$$

例えば、$n=100$とします。すると階級数（観測値をこれらそれぞれの階級数に割り振る）は[1]：

[1] この場合の階級数は大よその数と思うと気軽になります。これを使わなくてはならないことはないです。あくまで標本の観測値を整理して、見やすく、またそこから母集団の性質を知ることが大切ですからね。

$$階級数＝1+3.3x\log_{10}100=1+3.3x2=1+6.6=7.6\cong 8、$$ [2]

で、この大よその目安の階級数をどのように使うのかな？

Step 2: 階級の長さ（幅）を見つけます。
標本の観測値（ここでは観測個数$n=75$としょう）の中から最小値と最大値を見つけてください。例えば、観測値は整数で、最小値＝79、最大値＝１２９、としましょう。すると、すべての観測値が 79 から 129 に含まれることになります。

階級数はおおよそ８ときめていますので（なぜ？注２をみてください)、最大値と最小値の差を８で割ると、一つの階級（グループ）の長さ（もしくは幅）が出せます。この階級の幅を英語では, class interval　と言います。

階級の幅＝（最大値－最小値）÷階級数
Class interval = (the largest value – the smallest value)÷number of classes

階級の幅＝（最大値－最小値）÷階級数＝（129－79）÷8＝6.25≅６

Step 3:　最初の階級を決定します。
これまでに階級数が８、一つの階級の幅＝６、が分かっています。そこで最初（1 番目）の階級（グループ）の数字は、79 を含まないといけないです。

では、最初の階級の数値を 79 としますか？　NO!!　79 はなんとなくしっくりしないですね。やはり 75 から始まる階級（グループ）が最高と言う感じです。いざ、度数分布表の作成！

例１

階　級（以上～未満）			
75	～	81	←―― 最初の階級
81	～	87	←―― 二番目の階級

例 1 の表記はどうですか？　大切なことですが、度数分布表を出来る限り見

[2] では、$n=75$ならば、階級数は大よそいくつになりますか？８となりましたか？$n=100$も 8 だったから、あまり変わらない。

やすくする必要があります。　すると、見やすい階級となると、

例２

階　級（以上～未満）

75　～　80　←――　最初の階級

80　～　95　←――　二番目の階級

では､例 1 と例 2 ではどちらがいいのでしょうか？　見易さでは例 2 がよく、では例１の良いところは？？？

次にいってから考えましょう。

Step 4: 次に階級値を見つけます。
それぞれの階級の中心を**階級値**といいます。この階級値が度数分布表の大切な値となります。度数分布表では、階級値を x_i と表示します。ですから、度数分布表では、階級値は標本の観測値と同じ役割を果たします。すると、この階級値を整数、しかも出来る限り簡単な数値にする必要があります。

では、上記の例１と例 2 ではどちらが良いのでしょうか？

例１

階級（以上～未満）	x_i
75　～　81	
81　～　87	

例２

階級（以上～未満）	x_i
75　～　80	
80　～　85	

階級値は、それぞれの階級の下限と上限の値を足して、２で割って得ます。

階級値＝（下限の値　＋　上限の値）÷２

例えば、例 1 の最初の階級である 75～81 の場合は、

階級値＝（75＋81）÷２＝78

例２の最初の階級値は、77.5 となりますね（なぜ？）。

例１

階級（以上～未満）	x_i
75 ～ 81	83
81 ～ 87	89

例２

階　級（以上～未満）	x_i
75 ～ 80	77.5
80 ～ 85	82.5

どちらがいいでしょうか？　階級値x_iでみると、例 1 が見やすいですね。どちらも一長一短ありますね。では、階級の下限と上限の値も階級値もよくなるような、階級を決められるでしょうか？これは、各自が考えることです。あくまでも、度数分布表は見やくすることがベストです。

階　級	x_i
75 ～ 79	77
80 ～ 84	82

階　級(以上～未満)	x_i
77 ～ 83	80
83 ～ 87	85

このような例も考えられます。どれが一番見やすいかは、表を作り人のセンスによりますね。ただ、度数分布表で、階級の見やすさか、階級値の見やすさか、のいずれを選択しなければならない場合は、見やすい階級値を選ぶことを薦めます。

通常、階級の下限と上限の値を 10 や 100 などのようにして、階級の幅を 10 とすると、一般に度数分布表は非常に見やすくなります。

（ただの興味から、どのようなデータがそうなのか、これまでの例を使って、逆をしてみましょう。）
まず、最初にこれまでを整理します。

１．　標本の観測個数、それぞれの観測値、

２．　階級数＝$1+3.3x\log_{10} n$　(対数$\log_{10}$は常用対数)

３．　階級の幅＝（最大値－最小値）÷階級数

４．　階級値＝（下限の値＋上限の値）÷２

ここでの遊びは、

(1) 階級の下限値と上限値、そして階級値を出来る限り見やすい値にすること。
(2) 階級値の幅を 10 とすること。
(3) 階級の数を 10 とすること。

まず、上に整理された１～４を、逆に４から１への手順を踏めばよいと思います。

１．　（４）階級値＝（下限の値＋上限の値）÷２

２．　（３）階級の幅＝（最大値－最小値）÷階級数

３．　（２）階級数＝$1+3.3x\log_{10} n$　(対数$\log_{10}$は常用対数)

４．　（１）標本の観測個数、それぞれの観測値

想定として、データは前回と同様に最小値を 79 としましょう。

まず、

１．　最初の階級値は、80 がいいですね。

２．　すると、最初の階級の下限は 75、上限は 85 となります。階級の幅は 10 となります。

３．　階級数を 10 とすると決めていますから、

$10＝1+3.3x\log_{10} n$　(対数$\log_{10}$は常用対数)

$\log_{10} n=(10-1)\div 3.3=2.2727$

2.2727 の逆常用対数は、$n=187.3699$

つまり、標本個数は約 188 個となります。すると大よそ標本個数が 200 位だと、階級数が 10 位になるなと、考えられます。（注意、これが正しいとは言っていません。）

４．　さて、標本の最小値が 79、階級数が１０、階級の幅が１０、標本個数が 188（または 200 位）のときの、最大値は、どのくらいかな？または、最後の階級の下限と上限の値は何だろうか？

階級の幅＝（最大値－最小値）÷階級数

$10=(x_{largest}-79)\div 10$

$x_{l\arg est} = (10x10) + 79) = 179$、最大値は 179 となります。

ですから、最後の階級は

階級			x_i
175	～	185	180

ここでも分かると思いますが、最後の階級は必ず最大値を含まないといけません。

例えば、もし最大値が 179 ではなくて、１９６（適当ですが）ならば、階級数を 10 から 12 に増やして、

階級（以上～未満）			x_i
75	～	85	80
85	～	95	90
	・		
	・		
	・		
175	～	185	180
185	～	195	190
195	～	205	200

上の度数分布表はずいぶん見やすい表となりますね。また、階級幅を 20 とすると、階級の下限値と上限値が見やすくなります。ただし、この場合には、標本の観測値の最小値と最大値に大きな差が必要となります。

階級（以上～未満）			x_i
70	～	90	80
90	～	110	100

などが見やすいですね。[3]

[3] 階級幅は、上の表の例では、最初の階級の上限値が 90 であり、2 番目の階級の下限値が 90 ですので、どちらの 90 から 70 を引いても 20 となり、階級幅は 20 です。正式な階級の

[質問] 以下の度数分布表の階級幅はいくつでしょうか？

階　級	x_i
80 ～ 84	82
85 ～ 89	87
・	
・	

答えは、５です（なぜですか？注３を参照）。また、連続した二つの階級値の差が、階級幅となります（８７－８２＝５）。

随分わき道にそれましたが、度数分布表の作成にもどります。

Step 5: 標本の n 個の観測値をそれぞれの階級に割り振ります。そして、それぞれの階級に割り振られた観測値の数を数えて、階級値 x_i の横に、**度数** f_i（観測値の数）を記入します。ここでは、出来る限り見やすい階級値にします。最小値が 79、最大値が 127、とします。

度数分布表

階　級	x_i	f_i
78 ～ 82	80	
83 ～ 87	85	
・		
・		
128 ～ 132	130	

[質問] 教科書 45 ページの練習問題５をやってみましょう。また、累積度数と累積相対度数を加えましょう。

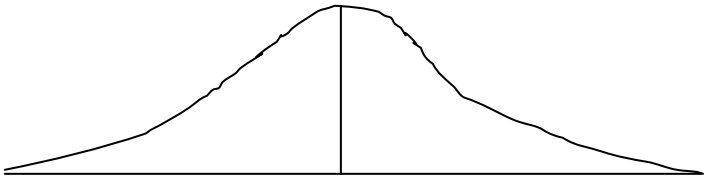
例えば、階級 $i=1$ から $i=l$ の度数を足した数を累積度数＝$\sum_{i=1}^{l} f_i$ と表します。また、累積相対度数は、累積度数を標本個数 n で割った値となります。

幅の求め方は、連続している二つの階級の、それぞれの上限値の差、もしくは下限値の差、となります。

1.3　グラフの形と名称

１．正規分布

平均値を中心に、左右対称なベル型をして分布を指します。この分布では、平均値、中央値、最頻値は同じ値となる。



平均値（中央値、最頻値）

２．右に歪んでいる分布（正の歪み：being skewed to the right）

右側方向に長い尾をもつ分布。分布が一番高くなる *x* の値（観測値）を最頻値という。その右側に中央値、そして一番右に平均値がくる。

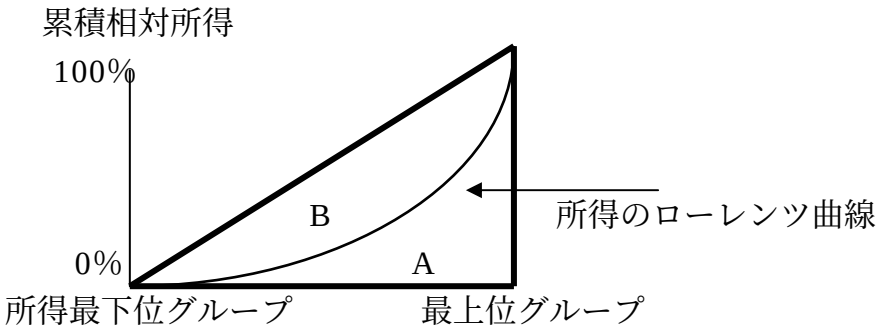
3．左に歪んでいる分布（負の歪み：being skewed to the left）

左方向に長い尾をもつ分布。分布が一番高くなる *x* の値（観測値）を最頻値という。その左側に中央値、そして一番左に平均値がくる。つまり、平均値、中央値、最頻値の順に並ぶ。

1.4　ローレンツ曲線 (Lorenz curve)

相対順位（X 軸）と累積相対分配（Y 軸）との関係を表す曲線。例えば、所得の低い準から、最下位の 20%のグループ、次の 20%のグループ、というように最上位の 20%まで 5 つのグループを X 軸に並べ、それらのグループまでの累積相対所得を Y 軸に取った関係を、所得のローレンツ曲線と言える。

もし、所得が平等に分配されていれば、最初の 20%のグループは全体の所得の 20%を分配されていて、次の 20%も全体の所得の 20%を分配されるので、これら 40%の人たちは所得の 40%を分配されていることになる。それぞれ 20%のグループが所得全体の 20%を所有することになるので、ローレンツ曲線は直線となる。



ジニ係数(Gini Coefficient)

不平等を表す指標。ジニ係数が０の場合は完全平等、１の場合は完全負平等を示す。ジニ係数は、上の図で表した面積 B と三角形の面積（A+B）との比で表すことが出来る。次のような式でも表すことができる。

ジニ係数＝１－{ローレンツ曲線下の多角形の面積/三角形の面積}
または、
ジニ係数＝２ｘ{相対順位と相対分布の積の総和}－{(n+1)/n}

以下は、クラスで示しませんが、機会があるときに紹介します。

1.5 発展したデータの代表値（各自、本を見てください。）

１． 幾何平均 (geometric mean)
２． 加重平均 (weighted mean)
３． 移動平均 (moving average or mean)
４． 加重移動平均 (weighted moving average)
５． 範囲 (range)
６． 変動係数 (coefficient variation)
７． 変数の標準化 (normalization of variable)

例えば、標本平均値 0、標本分散および標本標準偏差が１となるような変換は、$z=\left(\frac{1}{s}\right)(x_i-\bar{x})$ と表すことができる。z を標準正規変数(standard normal variate)といい、$z \sim N(0,1)$ と表すことができる。カッコの中の数値、０は平均値、１は分散をさす（もっとも、この場合は標準偏差の同じ数値である）。

偏差値$(50+10z)$は、標準化変換の一種で、平均が 50 点、標準偏差が 10 点となるように標準化された点数である。

1.6 物価指数（各自でチェック）

１． ラスパイレス指数：特定期（年）を基準として、その基準期（年）の購入量を一定とし、基準期と同じ数量を購入するための支出を基準期の価格で評価する。基準期の支出は分母とする。平たく言えば、基準期で買った購入量を当該期に購買すればいくらになるか、当該期の支出（分子）と基準期の支出（分母）との比である。

２． パーシェ指数：今期tと同じ数量Q_tをある期に買うとすれば、どれほどの

支出になるのか、今期の支出$\sum_{i=1}^{m} P_{ti}Q_{ti}$（分子）とある期 j の価格を使った支出$\sum_{i=1}^{m} P_{ji}Q_{ti}$（分母）との比である。

3． GNP デフレーター（パーシェ指数）

4． 海外物価指数

5． 購買力平価 (ppp, purchasing power parity)

1.7 2 変数データの整理（各自でチェック）

1． 標本共分散 (sample co-variance)

$$s_{xy}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

2． 標本相関係数 (sample correlation)

$$r_{xy} = \frac{s_{xy}^2}{\sqrt{s_x^2}\sqrt{s_y^2}}$$

ここでは、変数 X の分散を$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$と定義し、変数$Y$の分散も同様に定義される。