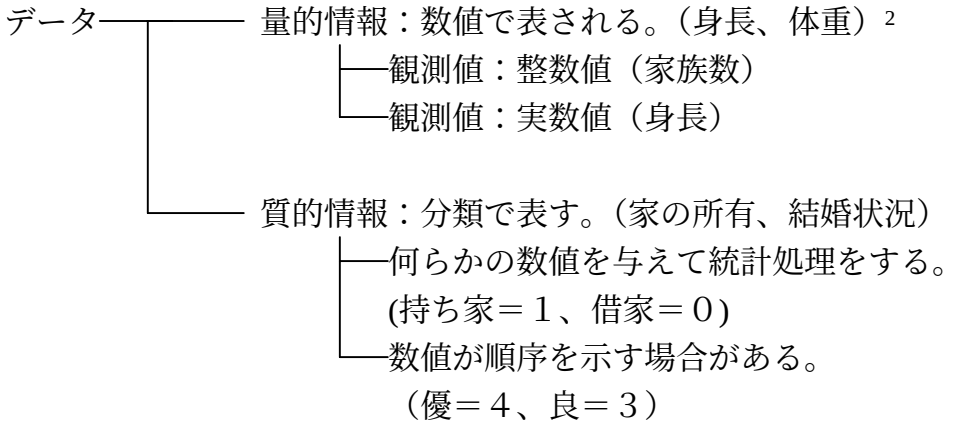
統計科学 (Statistics)：講義１

記述統計(descriptive statistics) [1]
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。

第一義統計：統計資料（統計処理を施した後の調査結果）を作成する目的で調査を行い、その結果を集計したもの。
第二義統計：もともとは統計資料の作成が目的ではない資料を集計して得た統計資料をさす。

１．データの整理

データ(data)
調査や実験から得られた情報（観測値）をまとめたもの。
時系列データ(time series data)：異なった時点での観測値からなるデータ。
クロスセクション(cross-section data)・データ：ある一時点で、異なった対象について観測値からなるデータ。
パネル･データ(panel data)：クロスセクション･データであるが、異なった時点（時系列データ）から得られた観測値からなるデータ。



[1] この講義は、東京大学教養学部統計学教室編『統計学入門』（東京大学出版会、1991 年）、p. 17、を参考にしている。
[2] 観測固体について、一つの観測値だけが与えられている場合を、１次元データ(1-dimensional data)と呼ぶ。一人の学生に対して一つに観測値（例えば、身長）だけをえる。２次元データ、二つの観測値、身長と体重、が得られた場合。３次元データ、身長、体重、性別。

1.1 データの代表値[3]

母集団：調査対象の全体 ◀― 全数調査

標本（データ）：一部の在学生を無作為に抽出する。

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：観測値（又は、観測値）が一定でない調査対象

観測値：$x_1, x_2, ..., x_n$.　$X = \{x_1, x_2, ..., x_n\}$. ◀―― 変数 X の計測された値の集まり。

観測値の集まりを集合とすると：　$S = \{x_1, x_2, ..., x_n\}$.　とも表せる。

母集団の総数（全数）： N

観測個数（標本数）：n

統計学では標本をつか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

平均(mean)：データの重心をさす。

1．算術平均 (arithmetic mean)[4]

観測値の総和を観測個数で割った値。

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

母集団：　$mean: \mu = \dfrac{\sum_{i=1}^{n} x_i}{N}$

Note: 母集団の平均値（Population mean）: μ（myu と読む）や m などの文字を使います。

---

[3] 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。

[4] 幾何平均(geometric mean)：　$x_c = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$，　$x_i > 0$. 成長率などの平均を求める時に使われる。例えば、ある期間の地価の年平均上昇率の平均を求める時など。

２．中央値 、メディアン(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。

度数分布表の中央値(median)の求め方：

$$Median = L + C\left(\frac{\frac{n}{2} - F}{f_m}\right)$$

L ：中央値のある階級の下限値(the lower limit of the median class)
C ：階級幅 (width of class interval, i.e., two successive lower (or upper) class boundaries)
n ：標本数(the number of observations in the data set (or sample)
F ：中央値のある階級までの累積度数(cumulative frequency up to but not including the median class's frequency)
f_m ：中央値のある度数(frequency of the median class)

３．最頻値 (mode)
最も頻繁に現れた値。分布の峰に対応する観測値。

例１：　$S = \{2,5,8,1,3,3,5,10,7,6\}$
mean= 5　; median= 5　; and mode=　3 と 5 .

例２：　$S = \{2,3,4,1,3,3,1,10,7,6\}$
平均値＝　　；中央値＝　　；最頻値＝　　.

度数分布表の最頻値の求め方：

$$Mode = L + C\left(\frac{d_1}{d_1 + d_2}\right)$$

d_1 ：frequency of the modal class minus the frequency of the previous class.

d_2 ：frequency of the modal class minus the frequency of the following class.

Note: 平均値（mean）は異常値に左右される、中央値や最頻値はこの限りでない。

データの広がりを示す代表値

観測値の散らばりの尺度[5]

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、不偏推定量（unbiased estimator）であるから。つまり、標本値から母集団値を推定する場合に、より正確な統計量（公式）である。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。

$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$ とした場合に、以下のように分解できる。

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

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

２．標本標準偏差 (sample standard deviation)

標本分散の平方根

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

[5] 散らばりの尺度のうち、観測値の最大値と最小値の差を範囲（レンジ：range）という。

３．全体（全標本）の母集団分散 (population variance)

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

４．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（平均値からのばらつき）を示す重要な統計量（またはパラメータ）である。

変動係数(coefficient of variation)

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

平均を考慮した上で散らばりを相対的に比較する指標。
例えば、アメリカの世帯所得分布と日本の世帯所得分布の比較など。つまり、世帯所得の平均値であるドルと円との直接比較はできないので、変動係数を用いて、どちらの国がより所得分布に偏りがあるのか調べることができる。

$C.V.(US) > C.V.(Japan)$　ならば、アメリカの世帯所得分布は日本のそれよりもより不平等となっている（偏りがある）といえる。

チェビシェフの不等式

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$\frac{1}{k^2}$以下である。

例：$k = 2$ の場合は、$\frac{1}{2^2} = \frac{1}{4} = .25(= 25\%)$ 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。

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

４．**十分位点（十分位階級）：第 1 十分位点、第 2 十分位点、....**
１０％ごとの値

５．**百分位点（パーセンタイル：percentile）：第 1 十分位点、第 2 十分位点、....**
１％ごとの値

1.2 度数分布表[6]

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

[6] 頻度分布ともいう。

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ことが大切！）：

$$\text{階級数} = 1 + 3.3 \times \log_{10} n \quad (\text{対数} \log_{10} \text{は常用対数})^{7}$$

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[8]：

$$\text{階級数} = 1 + 3.3x \log_{10} 100 = 1 + 3.3x2 = 1 + 6.6 = 7.6 \cong 8 \text{、}^{9}$$

　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。

---

[7] または、階級数 $k \approx 1 + \left(\dfrac{\log_{10} n}{\log_{10} 2} \right)$.

[8] 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。

[9] 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。

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（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 とします。
まず、

１．　最初の階級値は、80 がいいです。

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、

$10 = 1 + 3.3 \times \log_{10} n$　(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

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

$x_{largest} = (10 \times 10) + 79 = 179$、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。[10]

[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

[10] 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。

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	階級値　x_i	度数 f_i
78 ～ 82	80	
83 ～ 87	85	
・		
・		
128 ～ 132	130	