

第3回入学前勉強会

青森大学ソフトウェア情報学部

2024年度入学生対象

タイトル

「統計学・データサイエンスの入り口
：ヒストグラムから確率分布へ」

概要：

数学I「データの分析」の度数分布表、ヒストグラム、平均値、標準偏差などを復習をした後、正規分布や中心極限定理を紹介します。

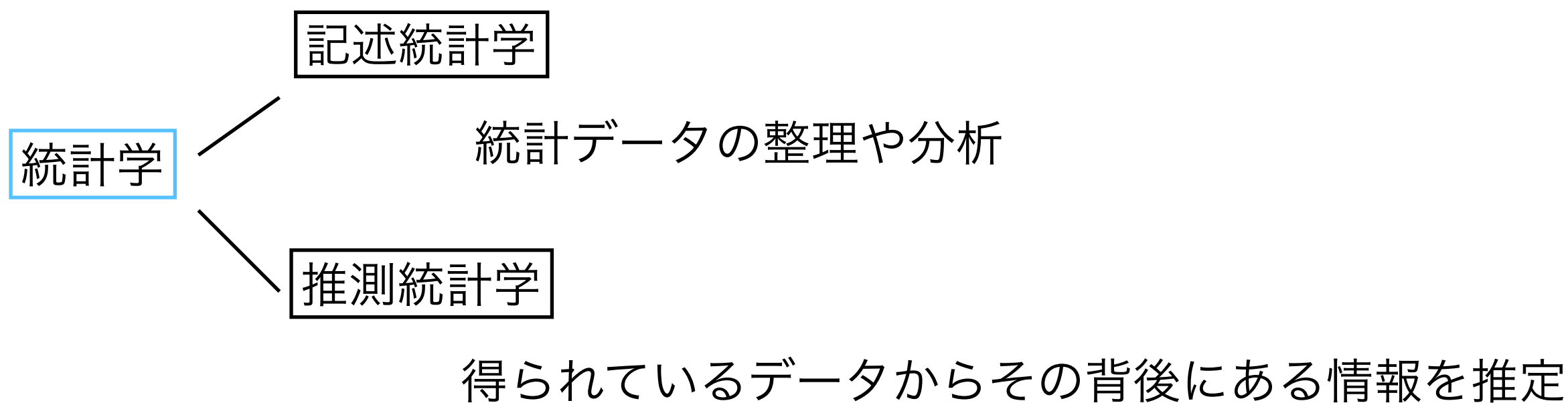
担当者：黒田茂

- 統計学は、記述統計学と推測統計学に分けることができます。記述統計学は、統計データの整理や分析に関するものです。推測統計学は、得られているデータからその背後にある情報の推定に関するものです。まずは、記述統計学を学びましょう。以下のトピックは、数学I「データ分析」の範囲に含まれる事柄です。

今日、ビッグデータの収集・蓄積・分析やAIを活用することにより、我々の社会が抱える様々な問題の解決や新たな価値を創造が期待されています。

そのため、数理・データサイエンス・AIに関する知識・技能について、基礎事項を習得しておくことが重要になってきています。

統計学は、数理・データサイエンス・AIの全てについての基礎の一つになっており、近年その重要性がますます高くなっています。



まずは、記述統計学を学びましょう。

以下のトピックは、数学I「データ分析」の範囲に含まれる事柄です。

以下の表は、X組10人の生徒の身長を記したものです。165や163のようにデータを構成している1つ1つの要素を観測値（または測定値）といいます。

表1：X組の生徒の身長

生徒	身長(cm)
1	165
2	163
3	169
4	170
5	176
6	164
7	166
8	174
9	165
10	168

観測値をただ並べただけではデータの特徴はわかりません。データの特徴を表や図でとらえ、活用する方法の一部を学びます。

最小値、最大値、範囲

- はじめに、データの散らばりの様子（^{ぶんぷ}分布という）を簡単につかもう。

- 最小値、最大値
- 範囲 = 最大値 - 最小値

X組の生徒の身長の
データ

生徒	身長 (cm)
1	165
2	163
3	169
4	170
5	176
6	164
7	166
8	174
9	165
10	168

チェック問題 問1

表1のデータの分布の最小値、最大値、および範囲を求めよ

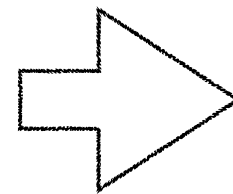
度数分布表

- 表2は表1のデータをもとに、160cmから175cmまでを5cmずつ区間に分け、各区間に入っている生徒の人数を調べて作成したものです。このような区間のことを「階級」、区間の幅を「階級の幅」、それぞれの階級に入っている観測値の個数を、その階級の「度数」といいます。またこのような階級と度数を示した表を「度数分布表」といいます。

上の
スライド
のデータ
だよ

X組の生徒の身長の
データ

生徒	身長 (cm)
1	165
2	163
3	169
4	170
5	176
6	164
7	166
8	174
9	165
10	168



X組の生徒の身長の
度数分布表

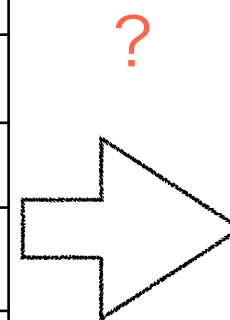
階級 (cm)	度数 (人)
以上 未満	
160 ~ 165	2
165 ~ 170	5
170 ~ 175	2
175 ~ 180	1
計	10

チェック問題 問2

左の表のデータは、ある中学校の1年A組の男子25人の体重を出席番号順に並べたものです。右の度数分布表を完成させよ。

1年A組男子の体重の
データ

生徒	体重(kg)	生徒	体重(kg)	生徒	体重(kg)
1	56	11	64	21	61
2	46	12	52	22	56
3	57	13	53	23	67
4	58	14	45	24	68
5	41	15	54	25	54
6	59	16	64		
7	50	17	47		
8	59	18	62		
9	51	19	58		
10	62	20	48		



度数分布表

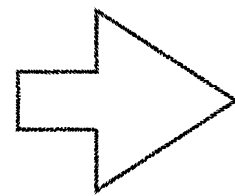
階級 (kg)	度数 (人)
以上 未満 40 ~ 45	
45 ~ 50	
50 ~ 55	
55 ~ 60	
60 ~ 65	
65 ~ 70	
計	

ヒストグラム

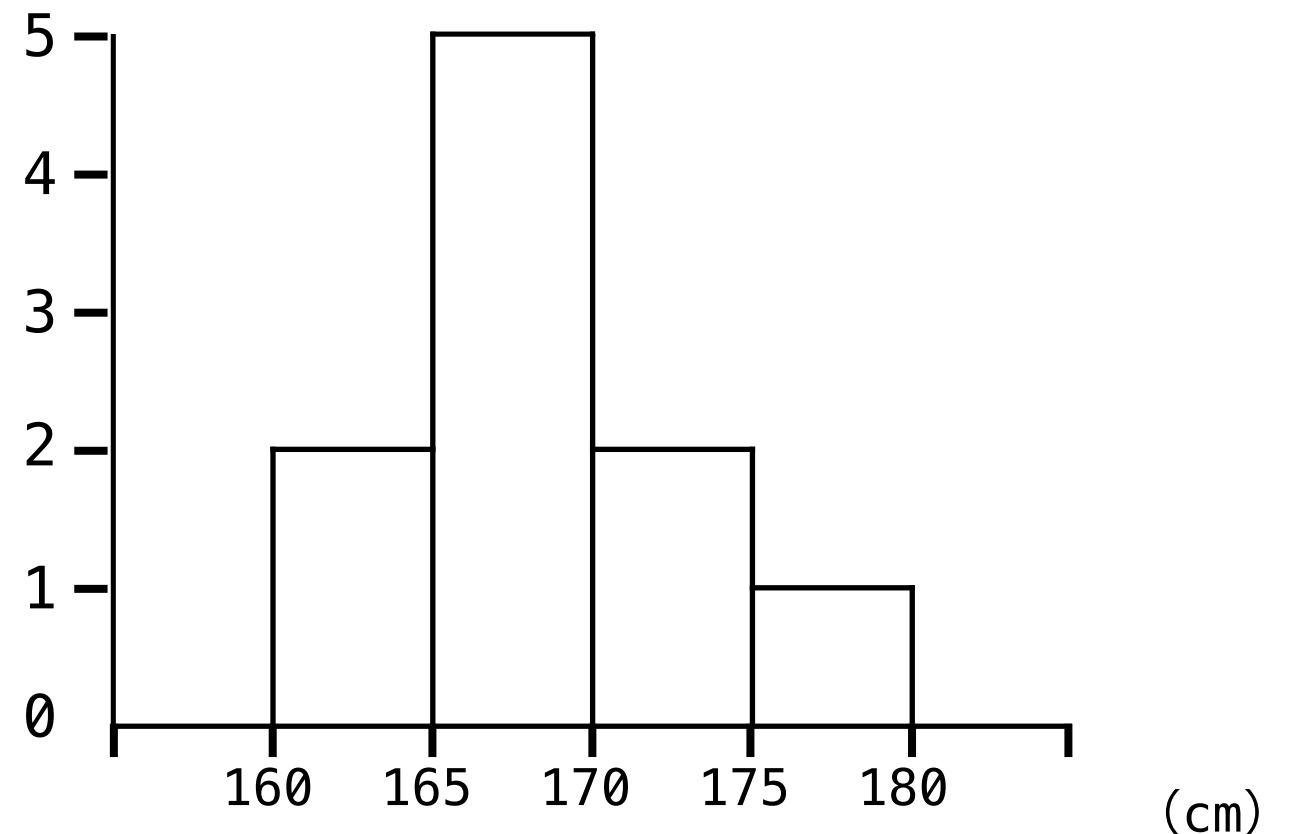
階級の幅を底辺とする長方形の面積が、その階級の度数に比例するように描いたグラフ

X組の生徒の身長
の度数分布表

階級 (cm)	度数 (人)
以上 未満 160~165	2
165 ~ 170	5
170 ~ 175	2
175 ~ 180	1
計	10



X組の生徒の身長
のヒストグラム
(人)

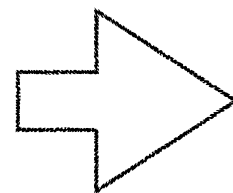


チェック問題 問3

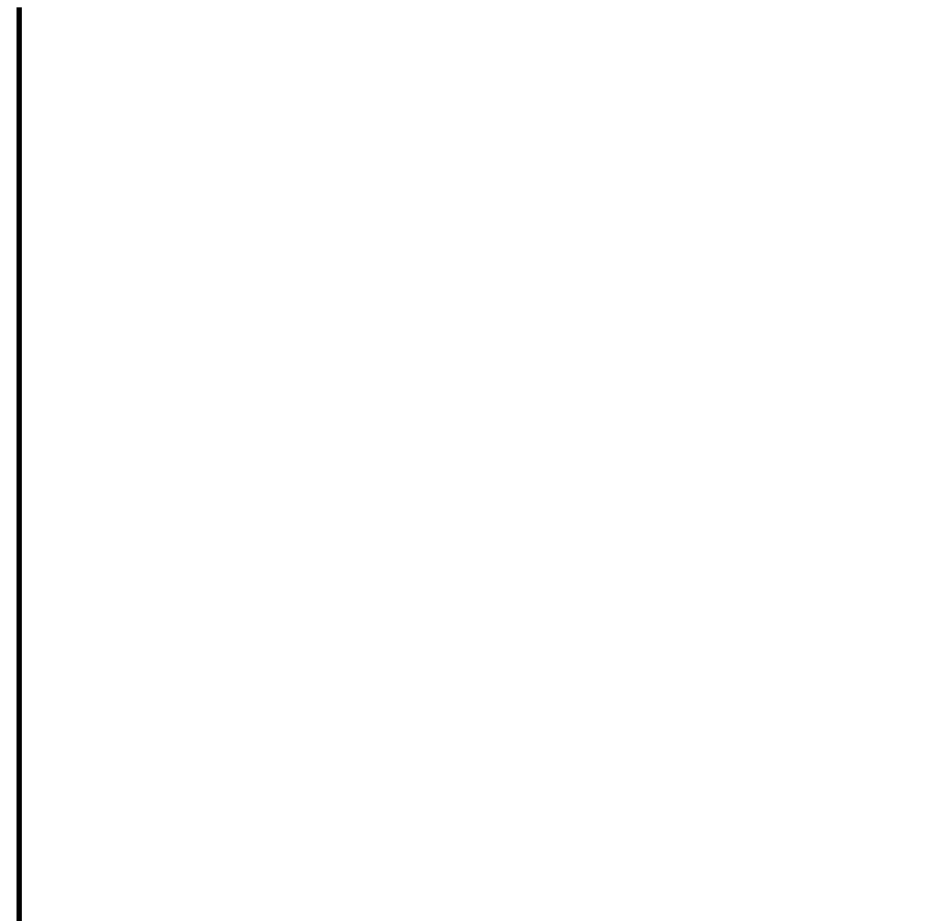
以下の度数分布表から、ヒストグラムを作成せよ。

度数分布表

階級 (kg)	度数 (人)
以上 未満	
40 ~ 45	1
45 ~ 50	4
50 ~ 55	6
55 ~ 60	7
60 ~ 65	5
65 ~ 70	2
計	25



ヒストグラム



・ 相対度数

A組の生徒の身長の数値分布表とB組の数値分布表を比較したい。

比べる際に気をつけること：

度数の合計（データサイズ）が同じであるか調べる。

下の例のようにデータサイズ（クラスの人数）が違っていたら、同じ階級の度数をそのまま比べても意味がない。

このようなときは、度数の代わりに度数の合計に対する割合、相対度数を用いる。

定義

$$(\text{相対度数}) = \frac{(\text{その階級の度数})}{(\text{度数の合計})}$$

A組の生徒の身長の数値分布表

階級 (cm)	度数 (人)
以上 未満 150 ~ 155	3
155 ~ 160	8
160 ~ 165	12
165 ~ 170	10
170 ~ 175	6
175 ~ 180	1
計	40

B組の生徒の身長の数値分布表

階級 (cm)	度数 (人)
以上 未満 150 ~ 155	5
155 ~ 160	8
160 ~ 165	11
165 ~ 170	13
170 ~ 175	6
175 ~ 180	5
計	48

・ 相対度数分布表

$$(\text{相対度数}) = \frac{(\text{その階級の度数})}{(\text{度数の合計})}$$

度数の代わりに、相対度数を書いた分布表

A組の生徒の身長の数、相対度数分布表

階級 (cm)	度数 (人)	相対度数
以上 未満 150 ~ 155	3	0.075
155 ~ 160	8	0.200
160 ~ 165	12	0.300
165 ~ 170	10	0.250
170 ~ 175	6	0.150
175 ~ 180	1	0.025
計	40	1.000

$$= \frac{3}{40}$$

$$= \frac{8}{40}$$

$$= \frac{12}{40}$$

(注意)
各階級の相対度数は、四捨五入の丸め誤差が入ることが多いため、必ずしも相対度数の合計が1になるとは限りません。

チェック問題 問4

A組の生徒の身長の相対度数分布表について、次の問に答えよ。

(1) 175cm 以上の生徒は全体の何%か。

(2) 165cm 以上の生徒は全体の何%か。

A組の生徒の身長の数値、相対度数分布表

階級 (cm)	度数 (人)	相対度数
以上 未満 150 ~ 155	3	0.075
155 ~ 160	8	0.200
160 ~ 165	12	0.300
165 ~ 170	10	0.250
170 ~ 175	6	0.150
175 ~ 180	1	0.025
計	40	1.000

チェック問題 問5

- (1) B組の生徒の身長の数値分布表から相対度数分布表を作成せよ。
- (2) 160cm 以上、170cm未満の生徒の相対度数はA組とB組ではどちらが大きいのか。

B組の生徒の身長の数値分布表

A組の生徒の身長の数値、相対度数分布表

階級 (cm)	度数 (人)	相対度数
以上 未満 150 ~ 155	3	0.075
155 ~ 160	8	0.200
160 ~ 165	12	0.300
165 ~ 170	10	0.250
170 ~ 175	6	0.150
175 ~ 180	1	0.025
計	40	1.000

階級 (cm)	度数 (人)	相対度数
以上 未満 150 ~ 155	5	
155 ~ 160	8	
160 ~ 165	11	
165 ~ 170	13	
170 ~ 175	6	
175 ~ 180	5	
計	48	

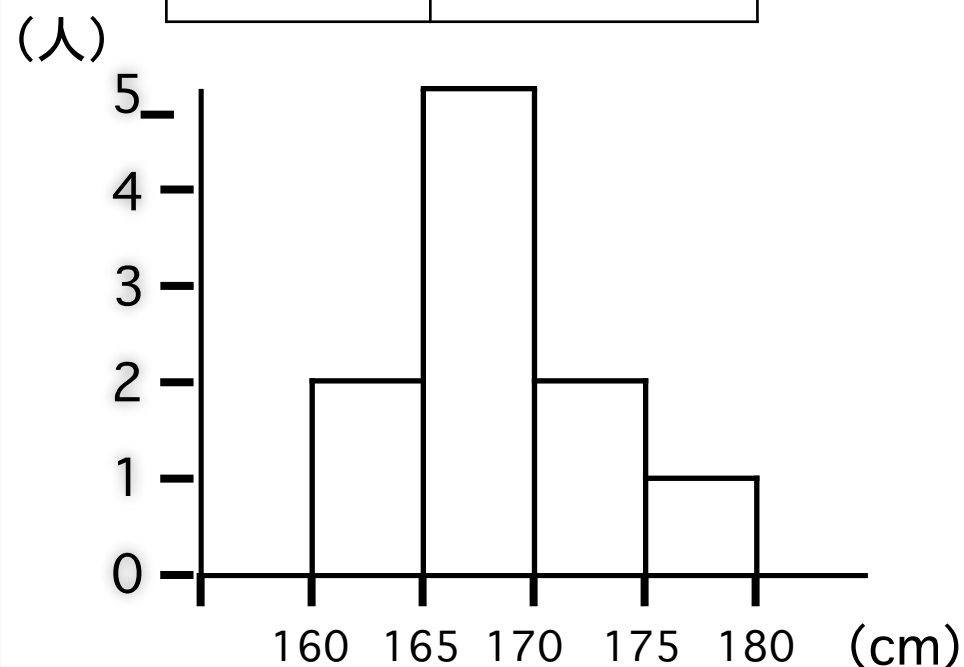
(補足) ヒストグラムは面積 \propto 度数

- ヒストグラム：階級の幅を底辺とする長方形の面積が、その階級の度数に比例するように描いたグラフ
- 棒グラフとは違う（棒グラフは棒の高さが度数に比例するように書いたグラフ。横軸は質的データ（種類で示されるもの）をとることが多い。）

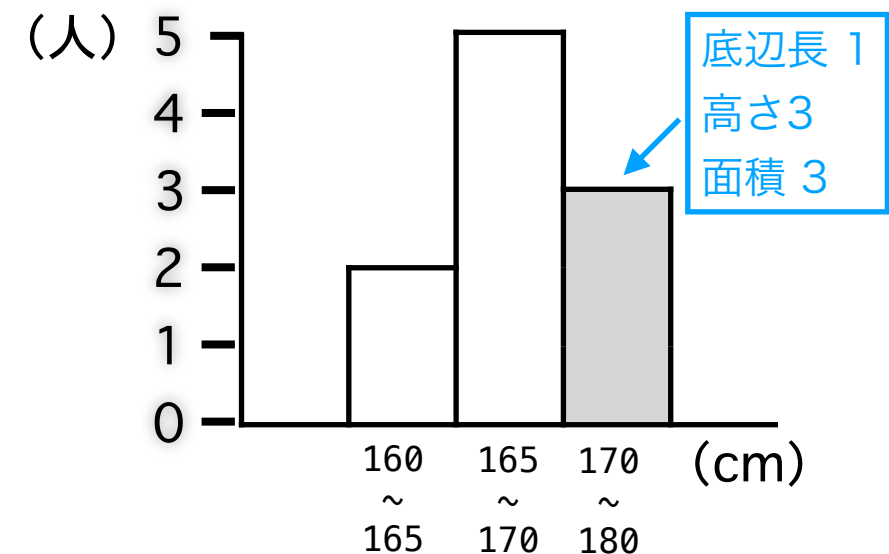
X組の生徒の身長の数値分布表

階級 (cm)	度数 (人)
以上 未満	
160 ~ 165	2
165 ~ 170	5
170 ~ 175	2
175 ~ 180	1
計	10

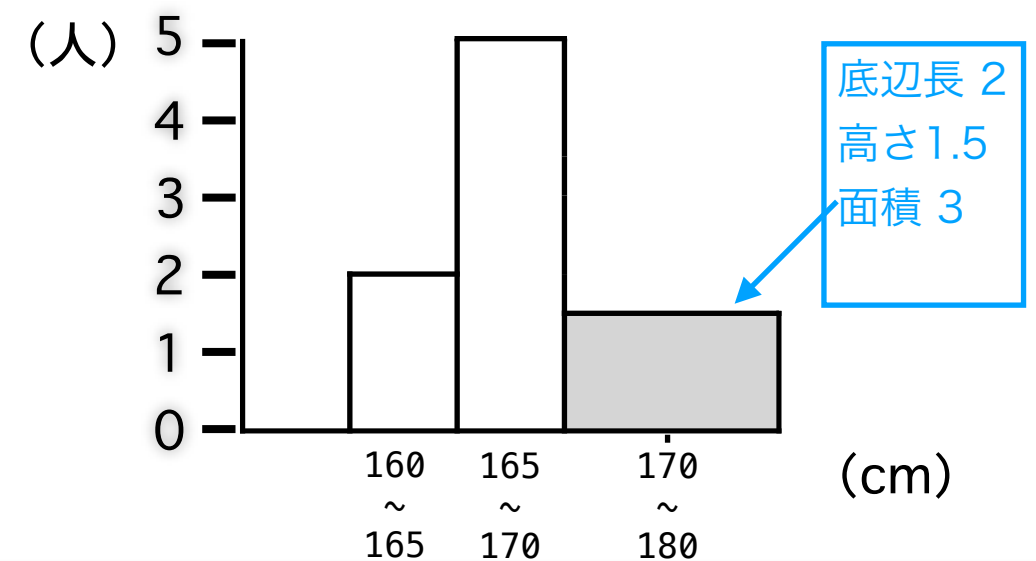
このデータについては一般的ではないと思うが、「階級の幅を一部変えて」ヒストグラムを作った場合の例（160~165, 165~170, 170~180）



パターン1

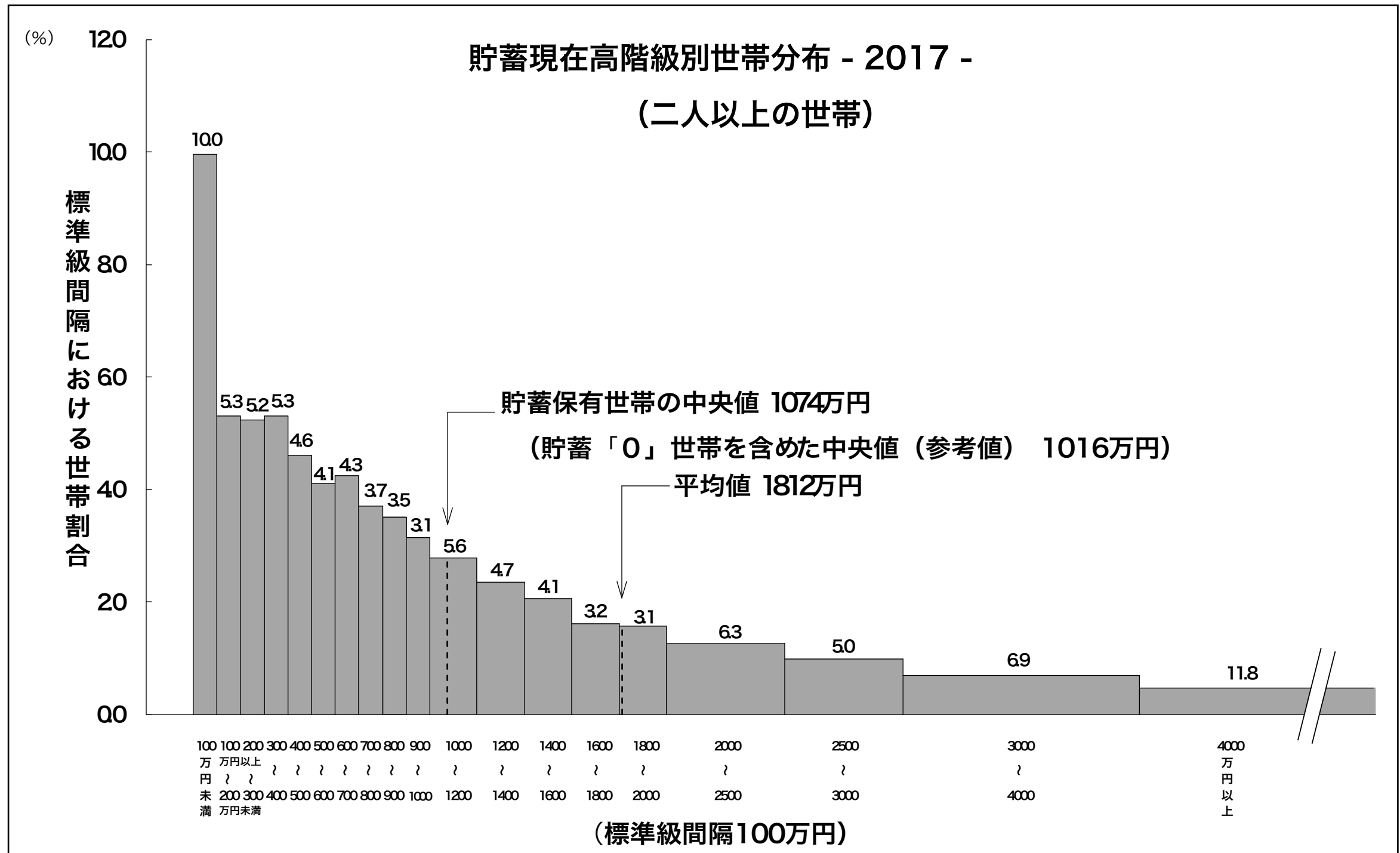


パターン2



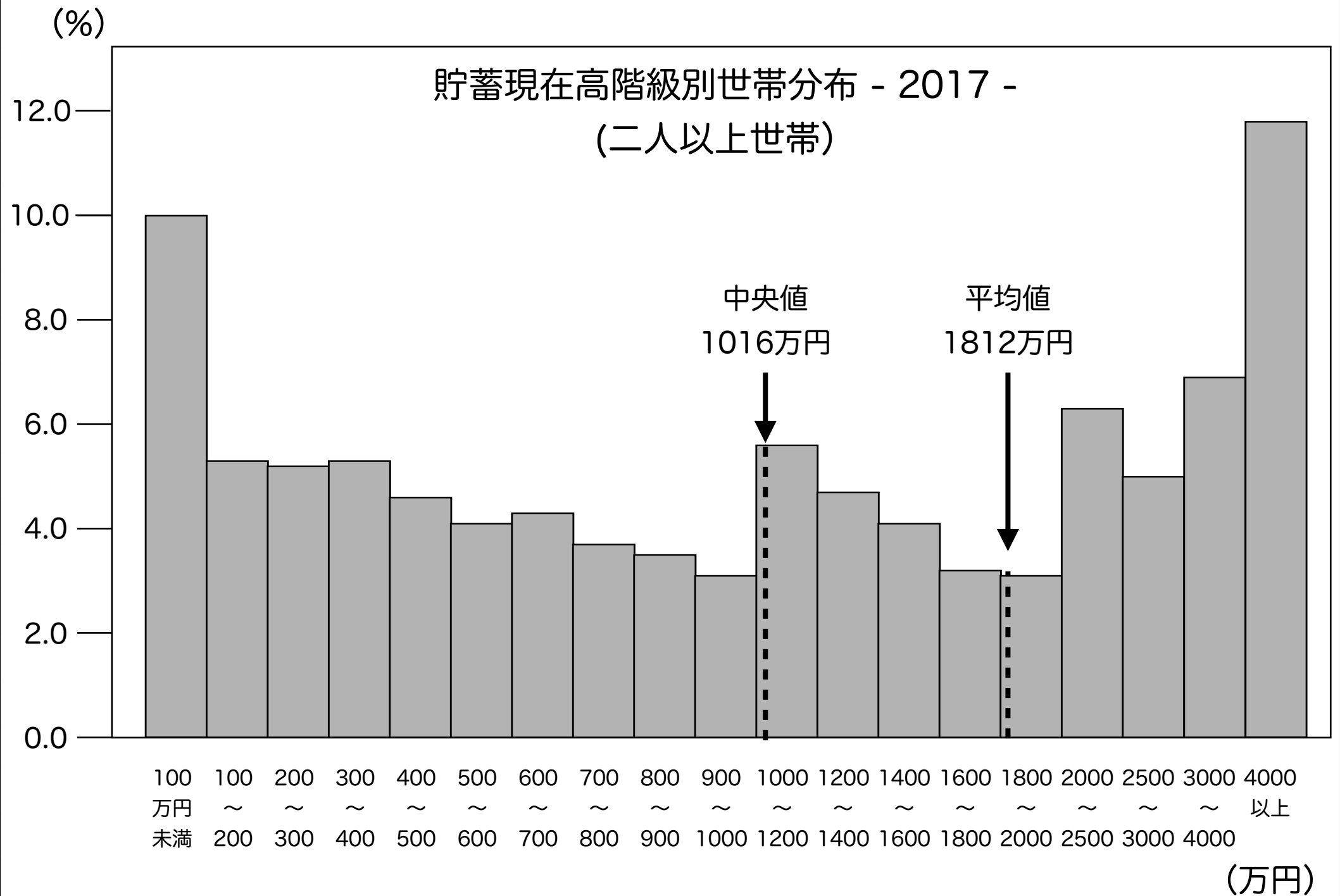
(参考)

総務省統計局の公開資料 (http://www.stat.go.jp/data/sav/sokuhou/nen/pdf/h29_gai2.pdf) より



底辺を揃えると見た目は大きく変わる

階級(万円)	度数
以上 未満	
0 ~ 100	10.0
100~200	5.3
200~300	5.2
300~400	5.3
400~500	4.6
500~600	4.1
600~700	4.3
700~800	3.7
800~900	3.5
900~1000	3.1
1000~1200	5.6
1200~1400	4.7
1400~1600	4.1
1600~1800	3.2
1800~2000	3.1
2000~2500	6.3
2500~3000	5.0
3000~4000	6.9
4000以上	11.8
計	99.8



代表値（平均値）

- ・ データの特徴を代表する1つの数値で表現することを考えます。
- ・ この一つの数値を代表値といいます。
- ・ 代表するとはいえ1つの値でデータの特徴を全て理解するのであるわけではないので、この後で説明する分布の散らばりの程度も含めて考える必要があります。
- ・ 複数のデータがあるとき、それらの代表値を比較することでデータに関する違いを大まかにすることができます。

例えば、試験の点数の平均値は、代表値の一つです。

前回の試験の平均点と今回の試験の平均点を比較して、これらの違いを考えることができます。

（よく使われる代表値の例として平均値の他に、中央値や最頻値などがあります。）

• 平均値

平均値は代表値の中で最も取り上げられるものです。

定義

$$(\text{平均値}) = \frac{(\text{観測値の合計})}{(\text{観測値の個数})}$$

(例1) データA : 0, 1, 2, 3, 4 の平均値は、

$$\frac{0 + 1 + 2 + 3 + 4}{5} = 2$$

(例2) データB : 1, 1, 1, 2, 2, 3, 3, 3, 3, 4, 4, 5 の平均値は、

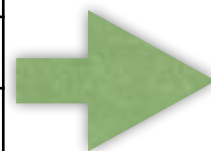
$$\frac{1 \times 3 + 2 \times 2 + 3 \times 4 + 4 \times 2 + 5 \times 1}{12} = \frac{32}{12} = \frac{8}{3} = 2.66\cdots \cong 2.7$$

(例3) 度数分布表から平均値を求める

階級は幅があるので、その真ん中の値をその階級の代表値として使います。これを **階級値** (階級の代表値) といいます。

A組の生徒の通学時間の
度数分布表

階級(分)	度数(人)
以上 未満 0 ~ 20	2
20 ~ 40	8
40 ~ 60	14
60 ~ 80	10
80 ~ 100	6
計	40



A組の生徒の通学時間の平均を求める表

階級(分)	階級値(分)	度数(人)	階級値x度数
以上 未満 0 ~ 20	10	2	20
20 ~ 40	30	8	240
40 ~ 60	50	14	700
60 ~ 80	70	10	700
80 ~ 100	90	6	540
計		40	2200

$$(\text{平均値}) = \frac{((\text{階級値}) \times (\text{度数})) \text{の合計}}{(\text{観測値の個数})} = \frac{2200}{40} = 55 \text{ (分)}$$

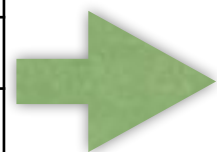
(例4) 度数分布表から平均値を求める(2)

$$(\text{平均値}) = \frac{((\text{階級値}) \times (\text{度数})) \text{の合計}}{(\text{観測値の個数})} \text{ は、 } (\text{相対度数}) = \frac{(\text{度数})}{(\text{観測値の個数})} \text{ より}$$

$$(\text{平均値}) = ((\text{階級値}) \times (\text{相対度数})) \text{の合計}$$

A組の生徒の通学時間の
度数分布表

階級(分)	度数(人)
以上 未満 0 ~ 20	2
20 ~ 40	8
40 ~ 60	14
60 ~ 80	10
80 ~ 100	6
計	40



A組の生徒の通学時間の平均を求める表

階級(分)	階級値(分)	度数(人)	相対度数
以上 未満 0 ~ 20	10	2	0.05
20 ~ 40	30	8	0.20
40 ~ 60	50	14	0.35
60 ~ 80	70	10	0.25
80 ~ 100	90	6	0.15
計		40	1

$$(\text{平均値}) = ((\text{階級値}) \times (\text{相対度数})) \text{の合計}$$

$$= 10 \times 0.05 + 30 \times 0.20 + 50 \times 0.35 + 70 \times 0.25 + 90 \times 0.15$$

$$= 55 \text{ (分)}$$

チェック問題 問6

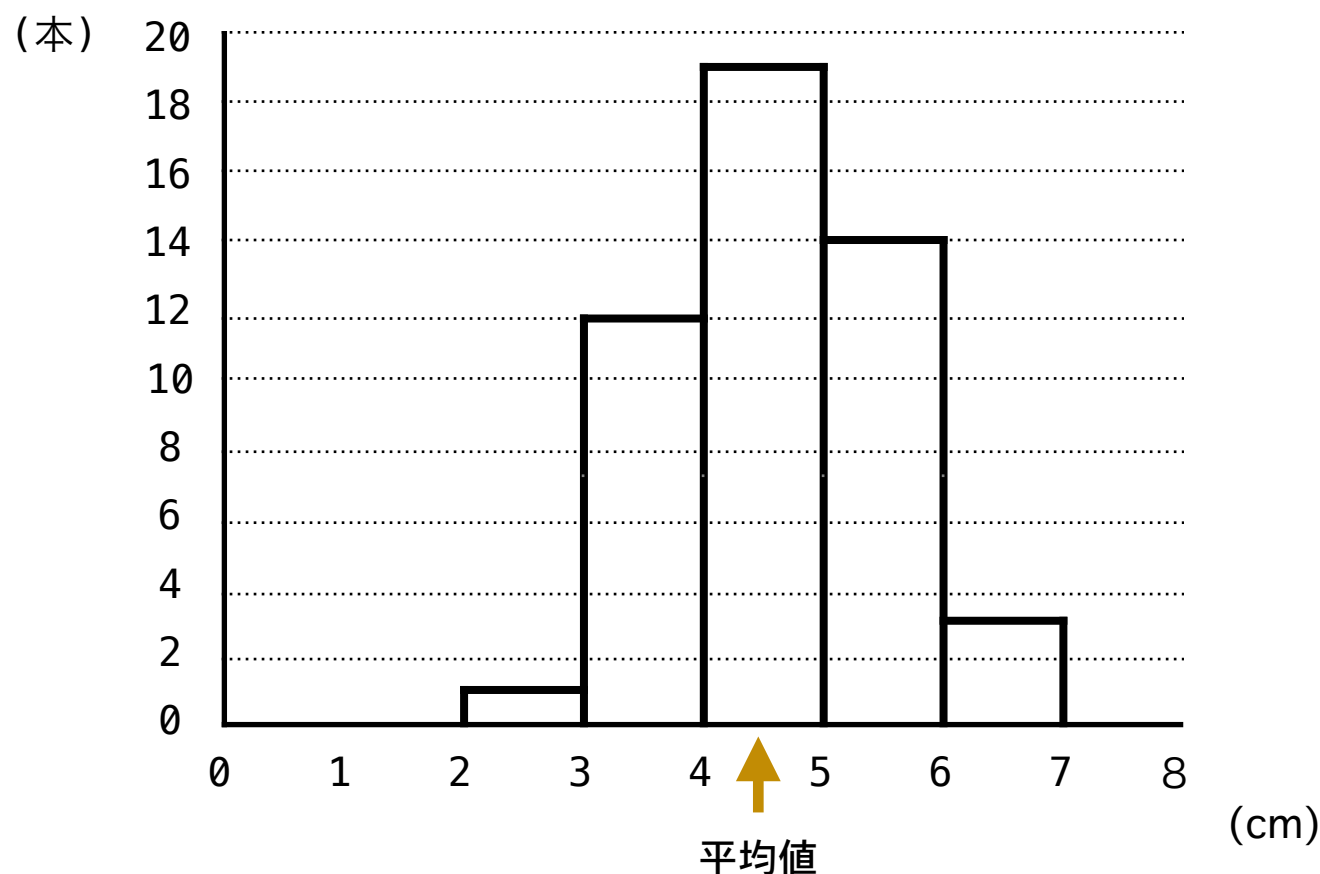
次の度数分布表は、B組40人の生徒の通学時間をまとめたものです。次の問いに答えよ。

- (1) 度数の合計、「階級値」の列および「相対度数」の列の空欄を埋めよ。
(2) 完成した表を利用して、（通学時間の）平均値を求めよ

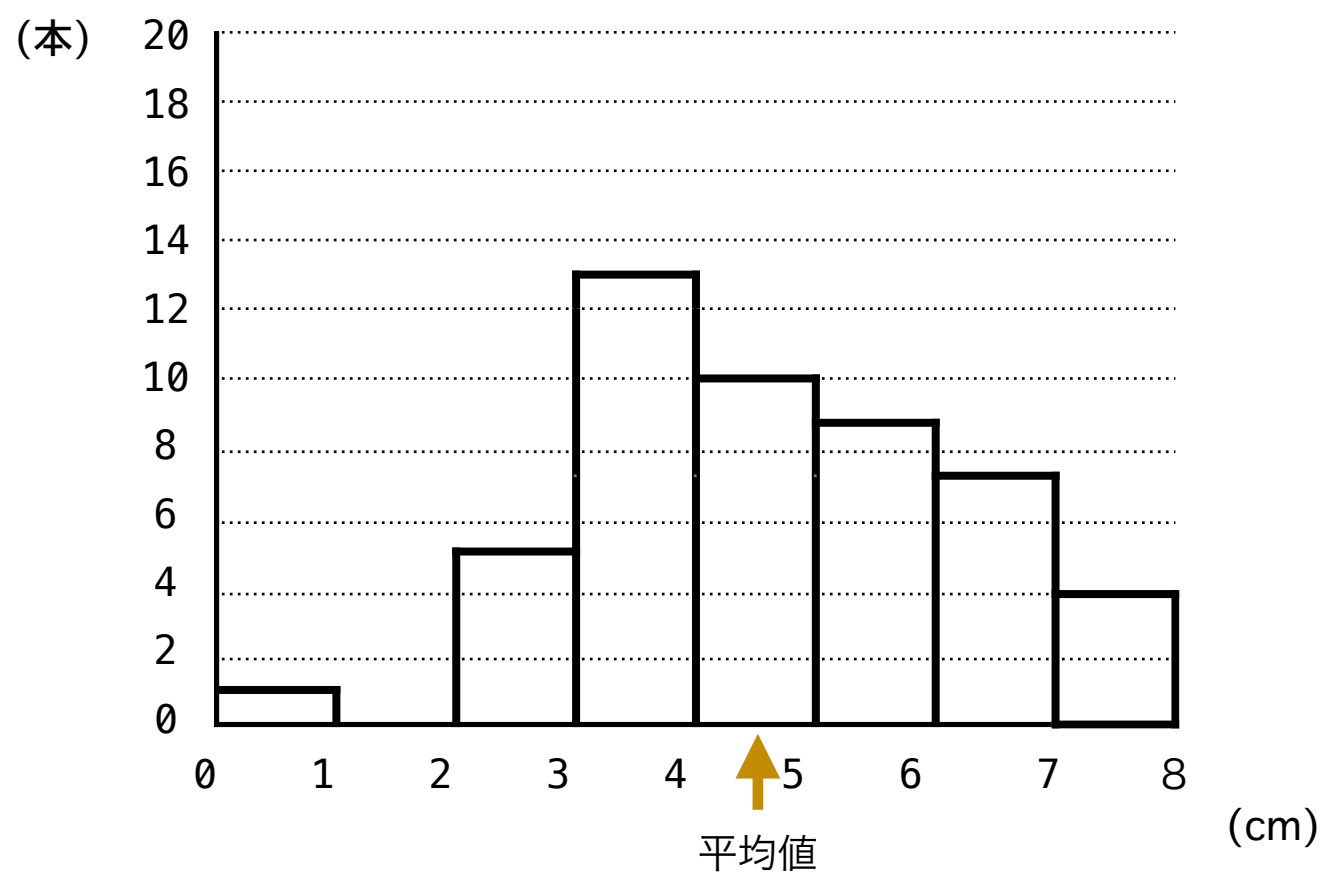
階級(分)	階級値(分)	度数(人)	相対度数
以上 未満 10 ~ 30		3	
30 ~ 50		12	
50 ~ 70		9	
70 ~ 90		10	
90 ~ 110		6	
計			

分布の散らばりの程度を表す指標 (分散と標準偏差)

- データの特徴を表すためには、代表値だけでは不十分な場合が多いです。例えば、下の例では、平均値はほとんど同じ（4.57と4.61）ですが、データの散らばりの程度は随分違うように見えます。
- そこで、データの散らばりの程度を数値で表現することを考えます。



A給食センターのポテト（49本）の長さのヒストグラム



B給食センターのポテト（49本）の長さのヒストグラム

- 「データが平均値を中心として、平均的にどの程度散らばっているのか？」という考え方を基本とします。

そこで、まず、各観測値について、平均からの離れ具合である**偏差**を次の式で定義します。

定義

$$(\text{偏差}) = (\text{観測値}) - (\text{平均値})$$

- 例 データ : 1, 2, 4, 5, 6, 7, 8, 8, 9, 10

$$\text{平均値} \quad \frac{1 + 2 + 4 + 5 + 6 + 7 + 8 + 8 + 9 + 10}{10} = \frac{60}{10} = 6$$

$$\text{偏差} \quad -5, -4, -2, -1, 0, 1, 2, 2, 3, 4$$

- ・ つぎに、この偏差をまとめて1つの数値にすることを考えます。

まず、思いつくのが偏差の和を計算することですが、

$$\text{偏差の和} = (-5) + (-4) + (-2) + (-1) + 0 + 1 + 2 + 2 + 3 + 4 = 0$$

となりマイナスとプラスが打ち消しあって偏差の和は常にゼロになります。

- そこで、打ち消しあい为了避免のために、偏差をそれぞれ2乗（平方ともいう）したものを利用します。

定義

$$(\text{偏差平方和}) = ((\text{観測値}) - (\text{平均値}))^2$$

- 例 データ : 1, 2, 4, 5, 6, 7, 8, 8, 9, 10

$$\text{平均値} \quad \frac{1 + 2 + 4 + 5 + 6 + 7 + 8 + 8 + 9 + 10}{10} = \frac{60}{10} = 6$$

$$\text{偏差} \quad -5, -4, -2, -1, 0, 1, 2, 2, 3, 4$$

$$\text{偏差平方和} \quad 25 + 16 + 4 + 1 + 0 + 1 + 4 + 4 + 9 + 16 = 80$$

- ・ この数値は、データの散らばりの程度を1つの数値で表していると言えますが、観測値の個数の影響を大きく受けています。観測値の個数が多いと単純にその分、値が大きくなります。
- ・ そこで、偏差平方和を観測値の個数で割って利用します。すなわち、偏差の平方の平均値を考えます。この値を**分散**といいます。

定義

$$(\text{分散}) = \frac{(\text{偏差平方和})}{(\text{観測値の個数})}$$

- ・ 例 データ：1, 2, 4, 5, 6, 7, 8, 8, 9, 10

- ・ 平均値 $\frac{1+2+4+5+6+7+8+8+9+10}{10} = \frac{60}{10} = 6$

- ・ 偏差 -5, -4, -2, -1, 0, 1, 2, 2, 3, 4

- ・ 偏差平方和 $25 + 16 + 4 + 1 + 0 + 1 + 4 + 4 + 9 + 16 = 80$

- ・ 分散 $\frac{80}{10} = 8$

- 次に、分散の単位について考えてみます。データの単位が (g) ならば、分散の単位は (g^2) , 単位が (cm) なら (cm^2) になります。これを元の単位と同じにするには、分散の正の平方根をとった値を考えます。この値を**標準偏差**といいます。

定義

$$(\text{標準偏差}) = \sqrt{(\text{分散})}$$

- 例 データ : 1, 2, 4, 5, 6, 7, 8, 8, 9, 10

$$\text{平均値} \quad \frac{1 + 2 + 4 + 5 + 6 + 7 + 8 + 8 + 9 + 10}{10} = \frac{60}{10} = 6$$

$$\text{分散} \quad \frac{80}{10} = 8$$

$$\text{標準偏差} \quad \sqrt{8} = 2\sqrt{2} \cong 2 \times 1.4 = 2.8$$

手順表の導入

・ 例 データ：1, 2, 4, 5, 6, 7, 8, 8, 9, 10

観測値 x	偏差 $x - \bar{x}$	偏差の平方 $(x - \bar{x})^2$
1	-5	25
2	-4	16
4	-2	4
5	-1	1
6	0	0
7	1	1
8	2	4
8	2	4
9	3	9
10	4	16
計 60		80

平均値 $\bar{x} = \frac{1 + 2 + 4 + 5 + 6 + 7 + 8 + 8 + 9 + 10}{10} = \frac{60}{10} = 6$

偏差：上の表の2列目

偏差の平方と平方和：上の表の3列目と3列目の一番下

分散 $s^2 = \frac{80}{10} = 8$

標準偏差 $s = \sqrt{8} = 2\sqrt{2} \cong 2 \times 1.4 = 2.8$

チェック問題 問7

次のデータについて以下に答えよ。

観測値 x	偏差 $x - \bar{x}$	偏差の平方 $(x - \bar{x})^2$
176		
170		
179		
188		
182		
計		

- (1) 平均値 $\bar{x} =$
- (2) 偏差の欄 (上の表の2列目) を埋めよ
- (3) 偏差の平方の欄 (上の表の3列目) を埋めよ
- (4) 偏差の平方和の欄 (上の表の3列目の一番下) を埋めよ。
- (5) 分散 $s^2 =$
- (6) 標準偏差 $s =$

さて、ここから少しだけ推測統計学のお話をします。

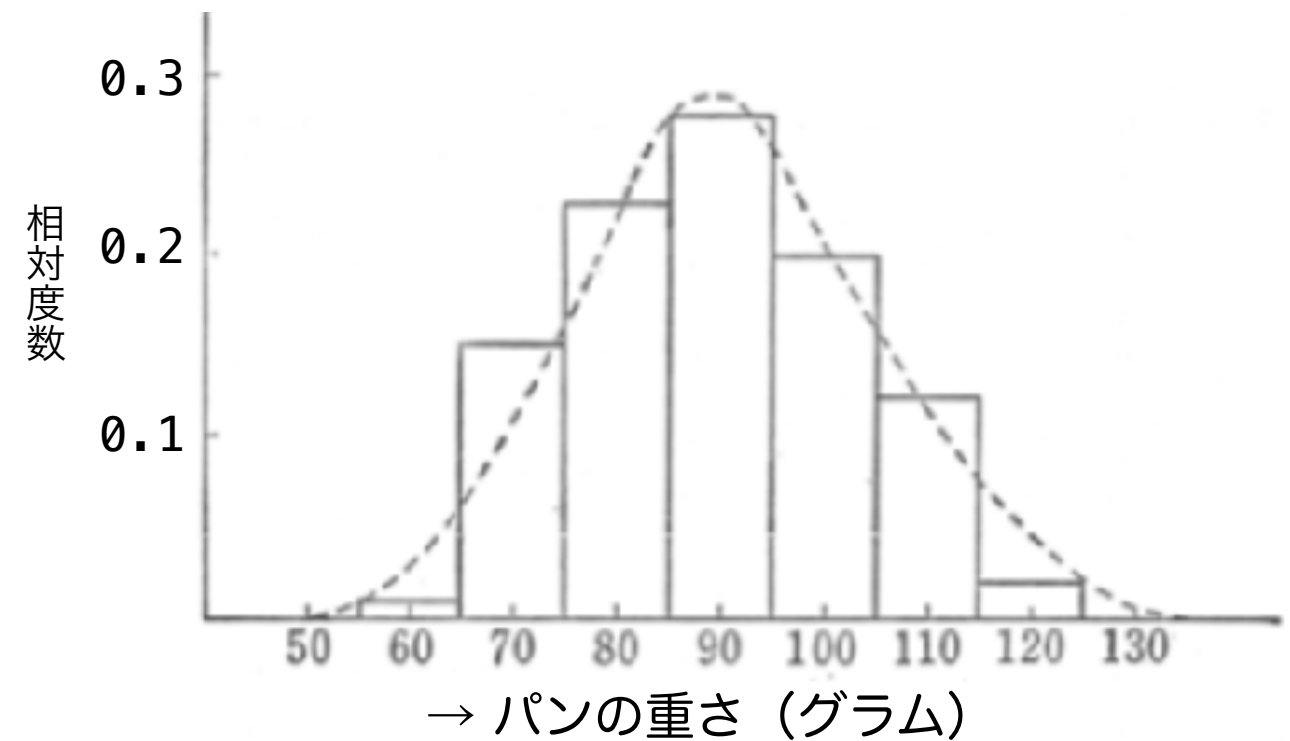
ここでのキーワードは、「正規分布」です。

正規分布と母集団

- あるパン屋で「1切れ100g」で販売されている
パンの重さを調べてみる

X氏が購入した100切れのパンの重さの度数分布表と相対度数ヒストグラム

パンの重さ [g]	個数	相対度数
以上 未満		
55~65	1	0.01
65~75	15	0.15
75~85	23	0.23
85~95	27	0.27
95~105	20	0.20
105~115	12	0.12
115~125	2	0.02
計	100	1



平均値を計算すると $89.4 \cong 90$ [g] だった

• 正規分布

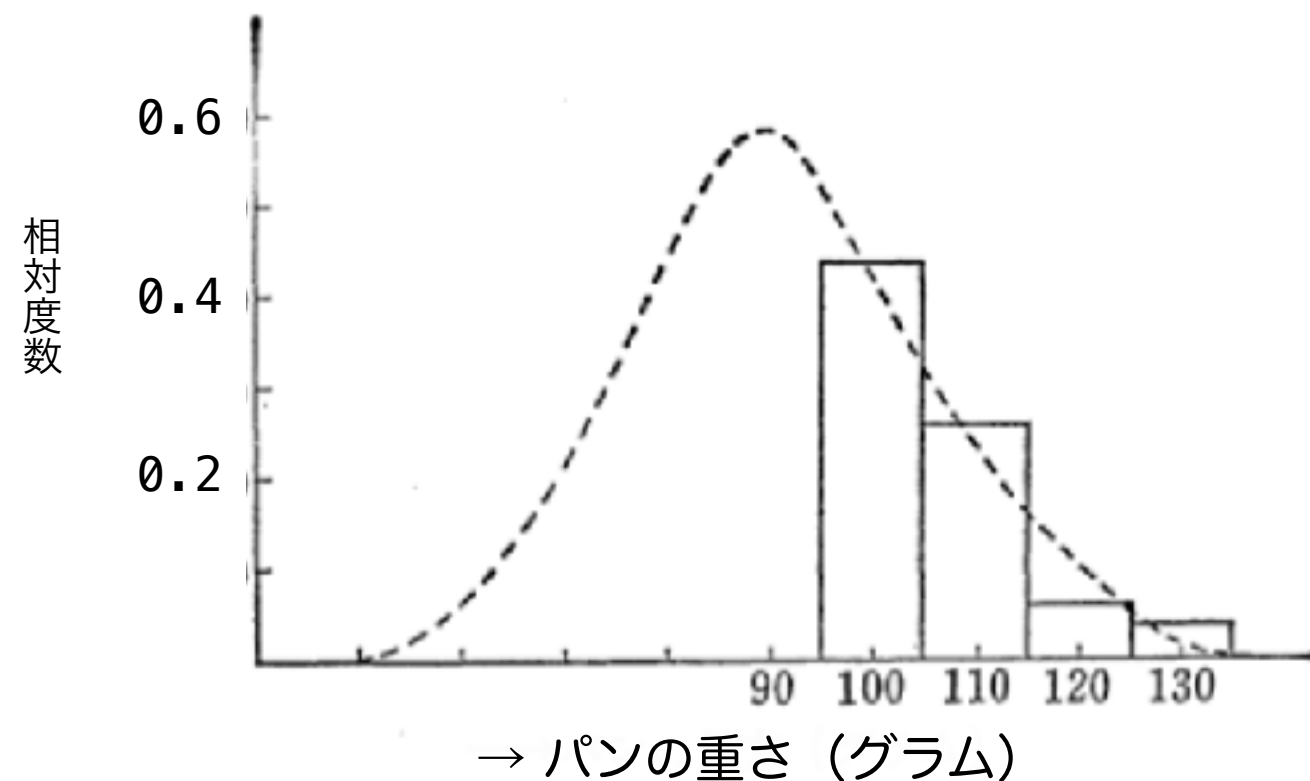
- ・ 一切れ90gのパンを作ろうと努力しても、すべて90gのパンを作るとは現実的に不可能で、90gを中心とする正規分布になる。（上のスライドの図の点線の曲線）
- ・ 正規分布をするものが多い測定値の例
 - ・ パンの重さのような製品の測定値、
 - ・ 人の身長などのような自然現象の測定値、
 - ・ 学生の成績などのような社会現象の測定値

（補足1）正規分布になる理由

- ・ パンの重さは、原料の性質、原料の配合、粉の練り方、発酵のできばえ、焼き方、パンの切り方など、多くの要因に影響される。
- ・ しかもこの要因の一つ一つは、完全にコントロールすることが不可能なので、それぞれ独立に変動する。
- ・ これらの変動の重なりが最終的にパンの重さの変動（バラツキ）になって現れてくる。
- ・ 一般に、このような場合には測定値が正規分布になることを数学的に証明することもできる。

- 別の相対度数ヒストグラムの例（正規分布でなかった場合）

同じパン屋さんで、別の人が購入したパン（100個）の重さをヒストグラムにしたところ次のようになった。



相対度数ヒストグラムが正規分布からずれている

→ 不自然

→ 工程のどこかに問題が発生している可能性がある

→ 工程管理に利用

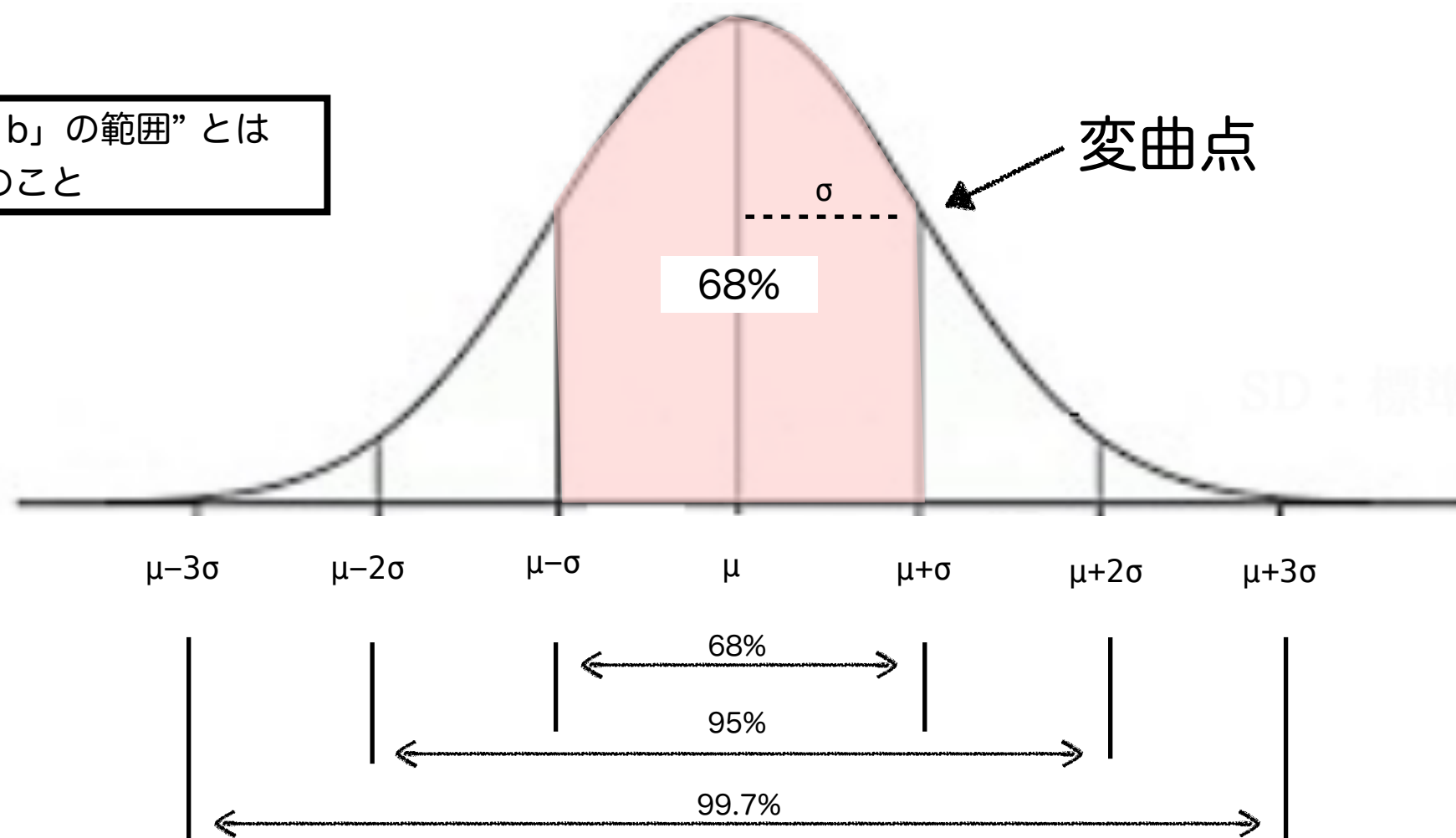
(補足2)

- 自然現象や社会現象のすべての分布が、正規分布になるわけではない。また、正規分布によく似た分布をしていても、厳密に検定してみると正規分布ではない場合も多い。たとえば、人間の体重の分布がその例であることが知られている。
- しかしながら、近似的に正規分布でありさえすれば、統計的な取り扱いに際して、実際には支障をきたさない。この点で、正規分布の価値は高いのである。

正規分布の特徴

- 左右対称な分布
- 平均 μ と標準偏差 σ によって分布の位置と形（中心と広がり具合）が決まる
- 「平均を中心にして±標準偏差」の範囲にあるデータが全体に占める割合は約68%
- 「平均を中心にして±標準偏差x2」の範囲にあるデータが全体に占める割合は約95%
- 「平均を中心にして±標準偏差x3」の範囲にあるデータが全体に占める割合は約99.7%

”「aを中心にして± b」の範囲”とは
a-b以上, a+b以下のこと



チェック問題 問8

正規分布において、

- (1) 平均以下のデータが全体に占める割合は何%か。
- (2) 「平均を中心にして \pm 標準偏差」の範囲の外にあるデータが全体に占める割合は何%か。
- (3) 平均+標準偏差 $\times 2$ 以上のデータが全体に占める割合は何%か。

チェック問題 問9

J国で国民全ての身長の調査を行なったところ、
平均 $\mu = 168$ cm、標準偏差 $\sigma = 7$ cm
であった。

J国のBMIの分布が正規分布であるとき、以下の空欄(a),(b)
を正しく埋めよ。

「J国民の95%の身長は

(a)

cm から

(b)

cm

の範囲に入ります。」

標本と母集団

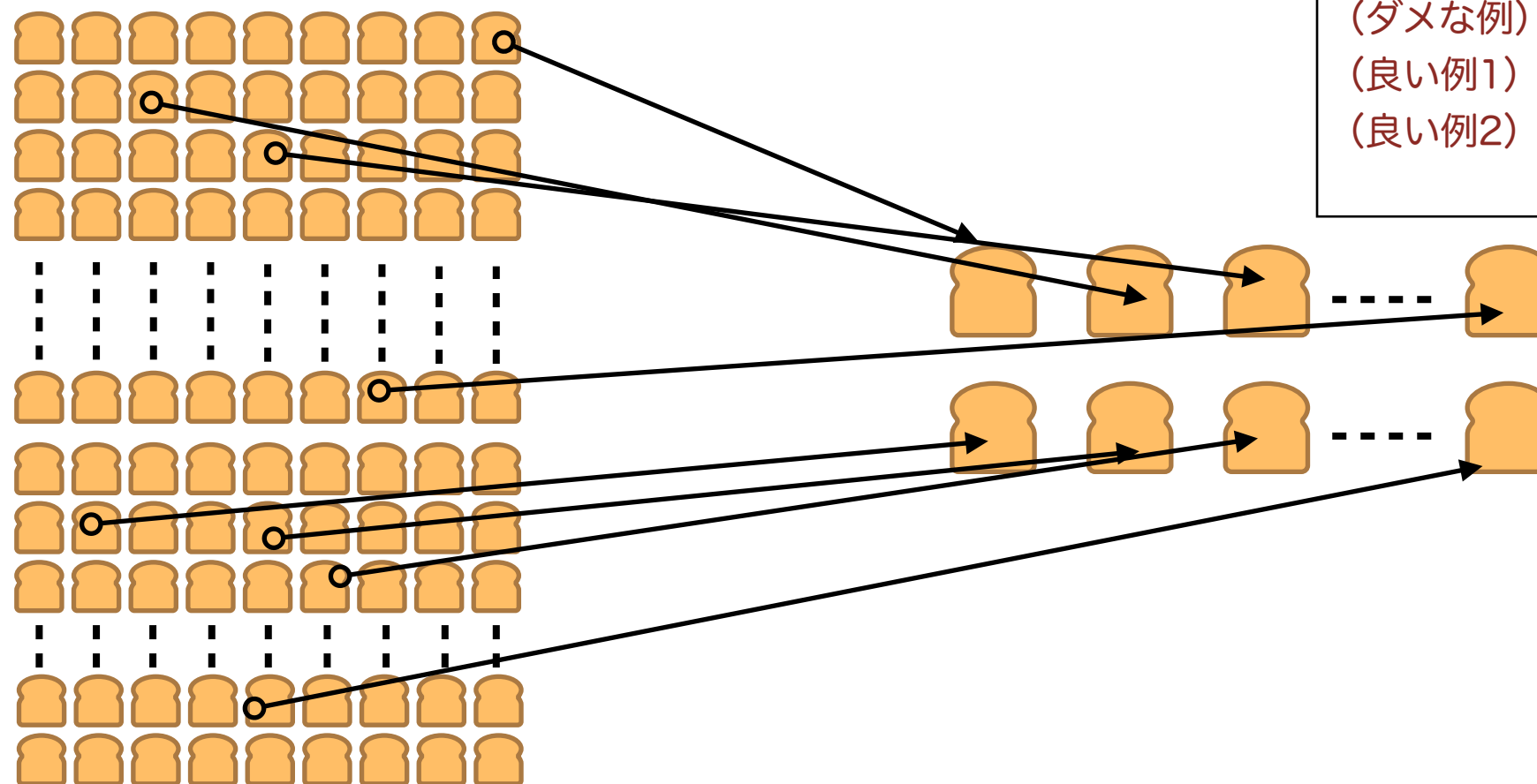
平均的な重さを推定する

- パン屋さんのパン一切れの（本当の）平均の重さを知りたい
- 全てのパンを調べるのは大変。
- ではどうすれば良いか。

母集団から標本を無作為抽出する。

無作為（ランダム）に抽出（サンプリング）した100切れのパンの重さから、販売しているパン全体の重さを推定する

- ・ 販売しているパン全て：母集団
- ・ 100切れのパン：標本（試料ともいう）



無作為抽出 = 味噌汁の味見

抽出する際、母集団の分布を反映するように気をつける。

（ダメな例）開店直後に全て買う

（良い例1）1時間ごとに抽出。

（良い例2）50個おきに抽出。

販売しているパン全て：母集団

抽出されたパン：標本

- ・母集団の平均、分散、標準偏差を、それぞれ特に、**母平均**、**母分散**、**母標準偏差**と呼び、

それぞれ、記号 μ （ミュー）、 σ^2 （シグマ二乗）、 σ （シグマ）で表す。

- ・理論的にわかっていること)

- ・ n 個の標本の観測値を用いた母平均の最も良い推定値

$$\text{標本平均 } \bar{x} = \frac{(\text{n個の標本の観測値 } x \text{ の総和})}{n}$$

これを仮に全ての標本の取り出し方について平均すると母平均と一致することが理論的にわかっている

●パン屋さんのパン1切れの重さの平均値の推定(1) (点推定)

パンの重さ [g]	個数	相対度数
以上 未満 55~65	1	0.01
65~75	15	0.15
75~85	23	0.23
85~95	27	0.27
95~105	20	0.20
105~115	12	0.12
115~125	2	0.02
計	100	1

$$\begin{aligned}
 & 1 \times 0.01 \\
 & + 15 \times 0.15 \\
 & + 23 \times 0.23 \\
 & + 27 \times 0.27 \\
 & + 20 \times 0.20 \\
 & + 12 \times 0.12 \\
 & + 2 \times 0.02 \\
 & = 89.4 \text{ (標本平均)}
 \end{aligned}$$

X氏が購入した100切れのパンの重さの標本平均は
89.4 [g]

→平均は 89.4 [g]と推定される

これで終わり？

標本平均が母平均の最も良い推定値だからといって、ぴったり一致するわけではない。標本を選び直すごとに毎回違ってくる。

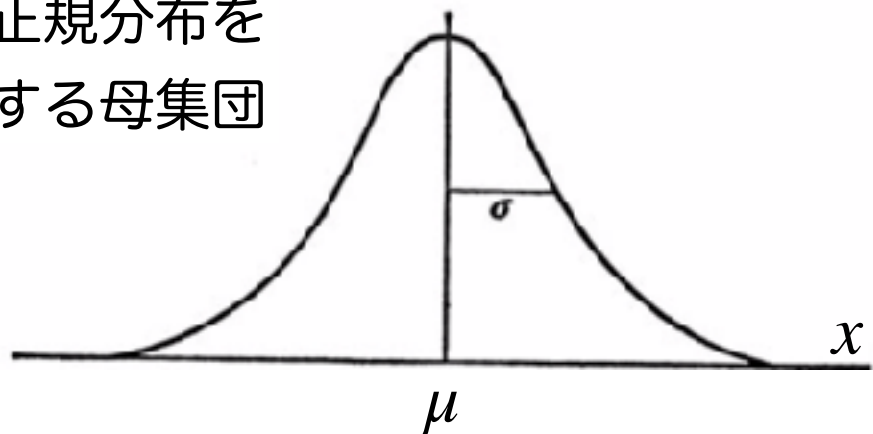
- ・ 例えば、1から10の数字からなる母集団から2個の標本を無作為抽出して平均を計算してみればすぐに納得：
 - ・ $\{1,3\} \rightarrow \text{標本平均} = 2$ 、 $\{5,2\} \rightarrow \text{標本平均} = 3.5$,

実は、母平均と標本平均との間に一般的に成り立つ関係について分かっていることがまだある。

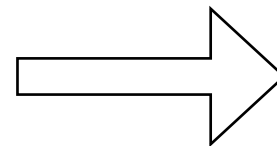
標本平均の分布に関する重要な定理

「母平均値が μ 、母標準偏差が σ の母集団から抽出した n 個の標本の標本平均 \bar{x} の分布は、 n を大きくするにつれて、平均が μ 、標準偏差が $\frac{\sigma}{\sqrt{n}}$ の正規分布に近づく。」

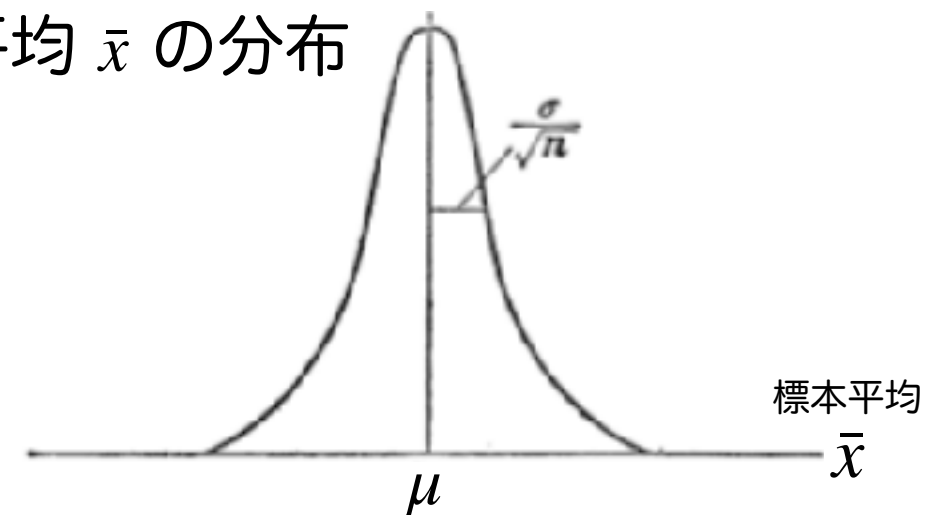
正規分布を
する母集団



n 個の x を抽出して
 \bar{x} を求める



標本平均 \bar{x} の分布



つまり、たくさんの標本を使って平均を計算した方がした方が母平均に近くなる（母平均に近い確率が高まる）

(例) 1から9までの数字が書いてあるサイコロを振った
時に出る数字の母集団

= 1から9までの数字が同じ割合で含まれている母集団

この場合、母平均、母分散、母標準偏差は簡単に計算できる。

サイコロの目	出現確率
1	1/9
2	1/9
3	1/9
4	1/9
5	1/9
6	1/9
7	1/9
8	1/9
9	1/9
計	1

$$\begin{aligned}\text{母平均} &= 1 \times (1/9) \\ &\quad + 2 \times (1/9) \\ &\quad + 3 \times (1/9) \\ &\quad + 4 \times (1/9) \\ &\quad + 5 \times (1/9) \\ &\quad + 6 \times (1/9) \\ &\quad + 7 \times (1/9) \\ &\quad + 8 \times (1/9) \\ &\quad + 9 \times (1/9) \\ &= (1+2+3+4+5+6+7+8+9)/9 \\ &= 45/9 \\ &= \underline{5}\end{aligned}$$

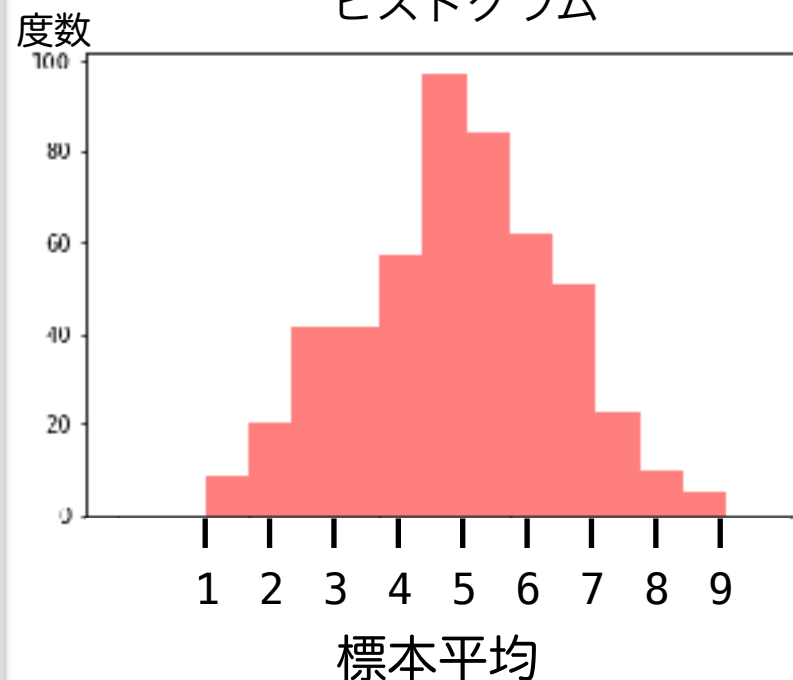
$$\begin{aligned}\text{母分散} &= (1-5)^2 \times (1/9) \\ &\quad + (2-5)^2 \times (1/9) \\ &\quad + (3-5)^2 \times (1/9) \\ &\quad + (4-5)^2 \times (1/9) \\ &\quad + (5-5)^2 \times (1/9) \\ &\quad + (6-5)^2 \times (1/9) \\ &\quad + (7-5)^2 \times (1/9) \\ &\quad + (8-5)^2 \times (1/9) \\ &\quad + (9-5)^2 \times (1/9) \\ &= 2 \times (4^2 + 3^2 + 2^2 + 1^2) \times (1/9) \\ &= 60/9 \cong \underline{6.6}\end{aligned}$$

$$\text{母標準偏差} = \sqrt{60/9} \cong \underline{2.6}$$

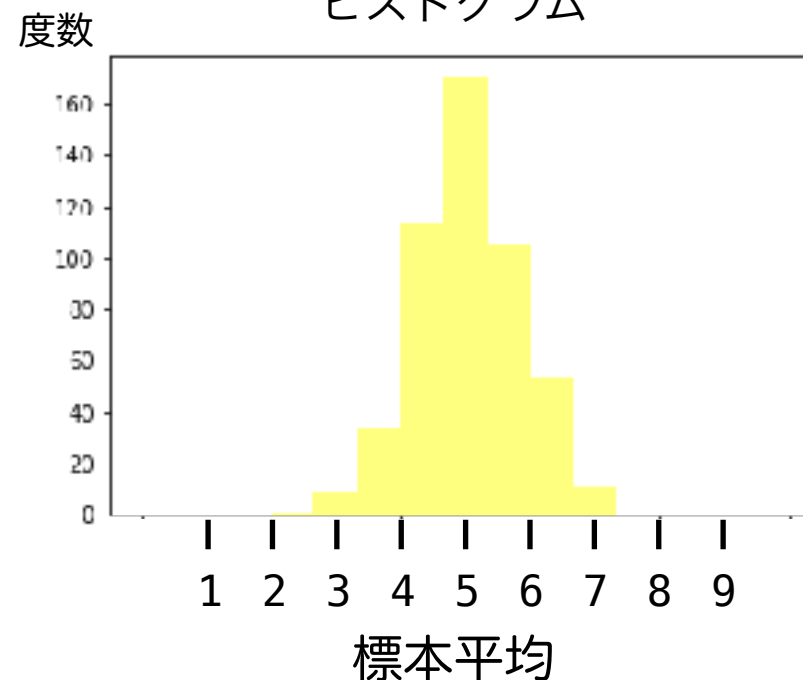
1から9までの数字をランダムにn個抽出し標本平均を計算することを500回繰り返した時のヒストグラムの例

これをn=3, 10, 100についてそれぞれ実行した

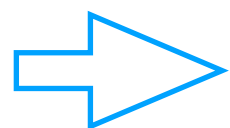
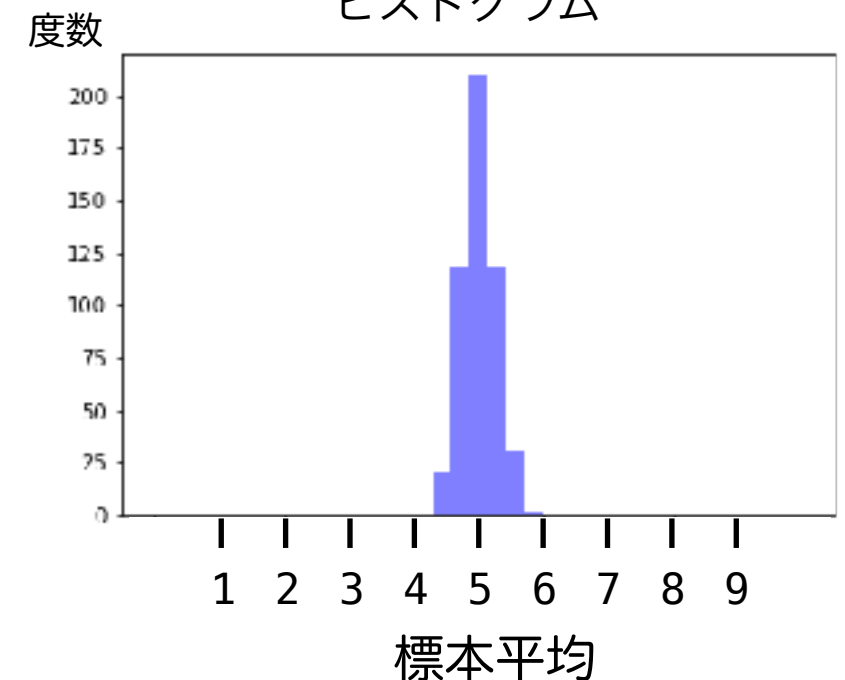
3個抽出して平均計算を
500回繰り返して作った
ヒストグラム



10個抽出して平均計算を
500回繰り返して作った
ヒストグラム



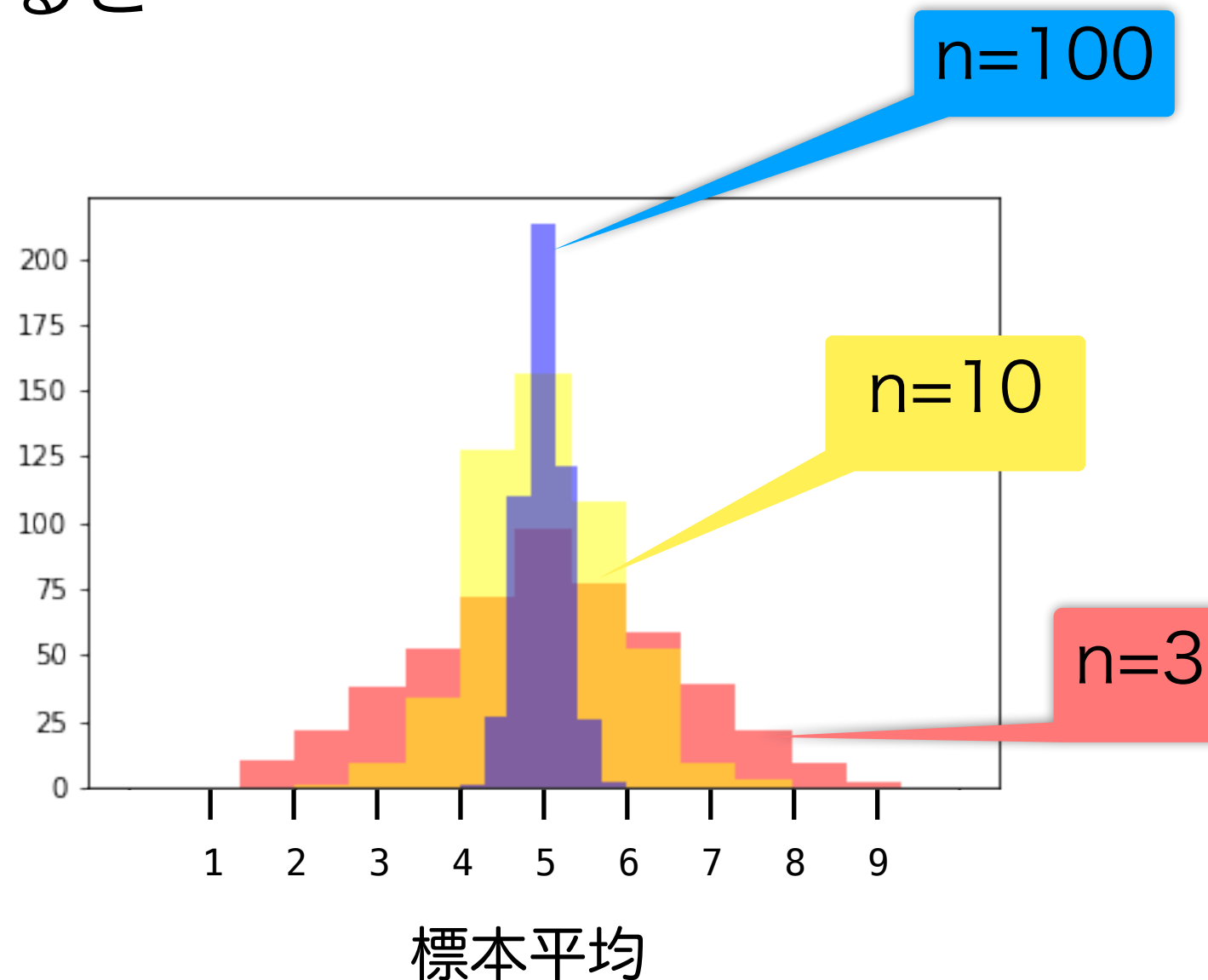
100個抽出して平均計算を
500回繰り返して作った
ヒストグラム



ヒストグラムは5を中心にした山形になっている。
平均を計算するために抽出する個数を増やすと、分布が真ん中に集中してきている。

重ねてみると

度数



平均5, 標準偏差 $2.6/\sqrt{n}$ の正規分布で近似できるとすると、n個抽出して計算した標本平均の

68%は、 $5 \pm \frac{2.6}{\sqrt{n}}$ の範囲にあるとか、 $5 + \frac{2.6}{\sqrt{n}} \times 3$ 以上の標本平均がでる確率は3%以下であるという見積もりがきる。

例えば、n=100のとき、500回の標本平均のうち概ね340個が $5 \pm 6.67/\sqrt{100} = 5 \pm 0.667$ の範囲、
の間にあり、標本平均が7以上になることは起こったとしても15個程度であると見積もれる。

上のスライドのヒストグラムを描くためのプログラムの例（Python言語を使用）

```
import numpy as np
import matplotlib.pyplot as plt

data = []
for i in range(500): # 以下を500回繰り返す
    x = np.random.randint(1, 10, 3) # 1以上9以下の整数を3個ランダムに取り出し
    # 平均を計算
    m = x.mean()
    data.append(m)

data2 = []
for i in range(500): # 以下を500回繰り返す
    x = np.random.randint(1, 10, 10) # 1以上9以下の整数を10個ランダムに取り出し
    # 平均を計算
    m = x.mean()
    data2.append(m)

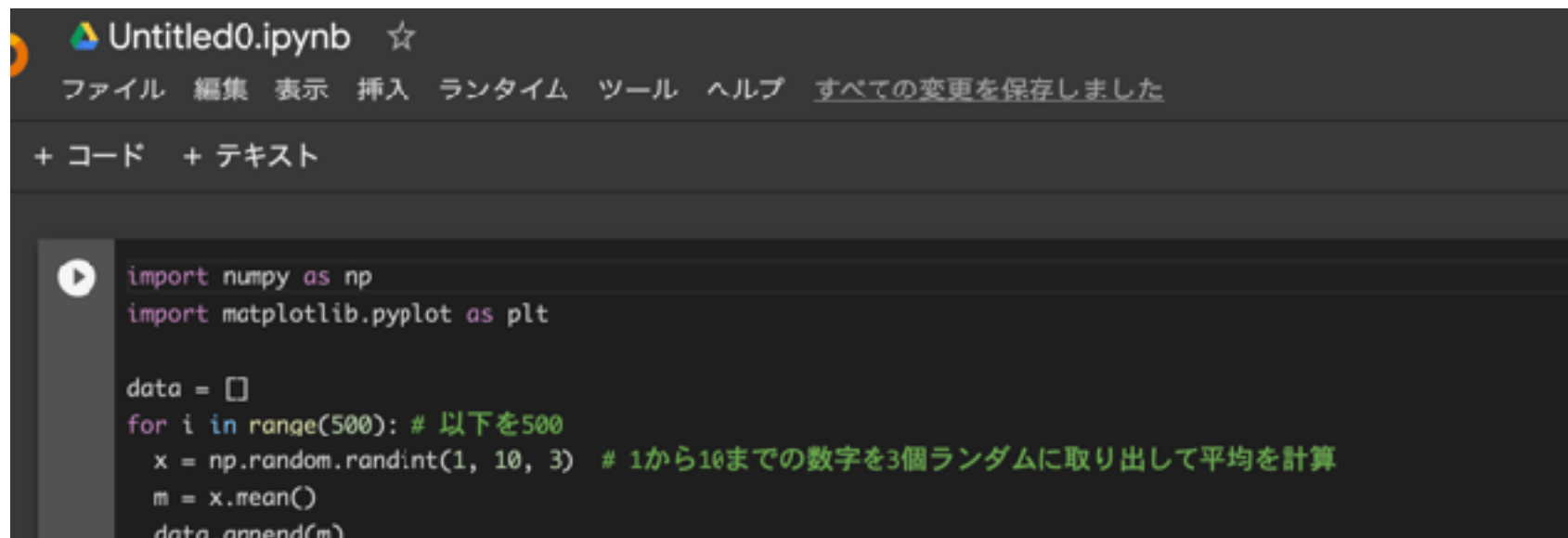
data3 = []
for i in range(500): # 以下を500回繰り返す
    x = np.random.randint(1, 10, 100) # 1以上9以下の整数を100個ランダムに取り出し
    # 平均を計算
    m = x.mean()
    data3.append(m)

# ヒストグラムを出力
fig = plt.figure()
ax = fig.add_subplot(1,1,1)
ax.hist(data, range=(0,10), color='red', alpha=0.5, bins=15)
ax.hist(data2, range=(0,10), color='yellow', alpha=0.5, bins=15)
ax.hist(data3, range=(0,10), color='blue', alpha=0.5, bins=35)
```

上のスライドのプログラムを自分で実行するには

- 簡単に使える実行環境 Google Colaboratory
 - ただし、Googleのアカウントが必要です。
- 使い方は、例えば以下のサイトを見て下さい。
 - <https://tracpath.com/works/development/google-colaboratory/>
- このサイトの「Google Colaboratoryで「Hello, World」してみる」できればOK.

上のスライドの プログラム をコピーして実行



```
import numpy as np
import matplotlib.pyplot as plt

data = []
for i in range(500): # 以下を500
    x = np.random.randint(1, 10, 3) # 1から10までの数字を3個ランダムに取り出して平均を計算
    m = x.mean()
    data.append(m)
```

パソコンを持っている人は是非やってみてね。