

データサイエンス入門B

第9回 平均・分散・標準偏差

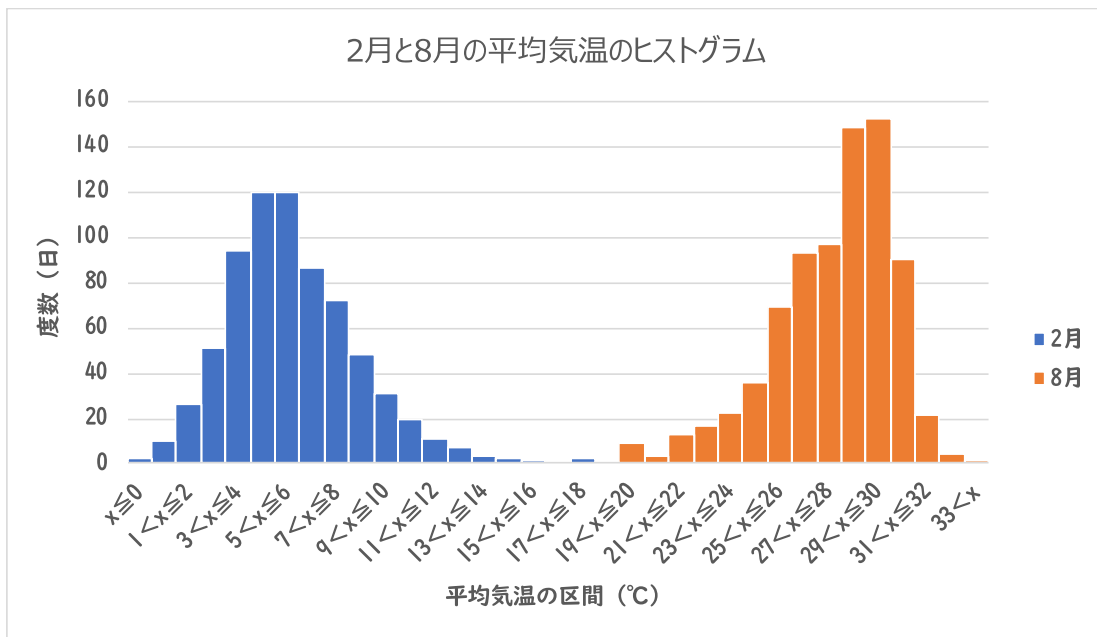
高田 美樹

目次

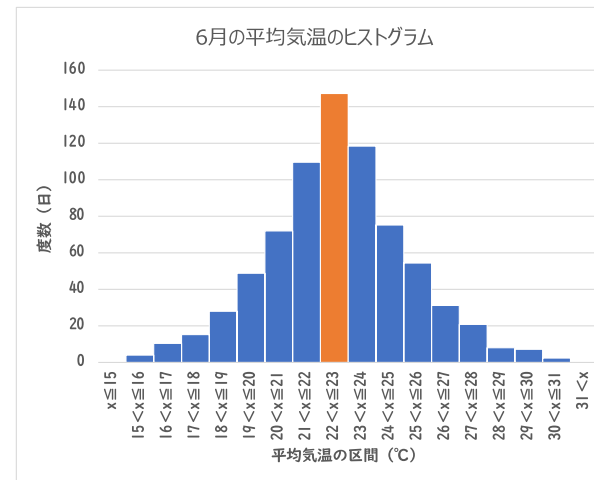
- ▶ 基本統計量（おさらい）
- ▶ 期待値と標準偏差
- ▶ 正規分布
- ▶ zスコアと偏差値
- ▶ 変動係数（cv）
- ▶ 尖度と歪度
- ▶ 記述統計と推測統計
- ▶ 標本分布
- ▶ 点推定
- ▶ 区間推定

前回のおさらい

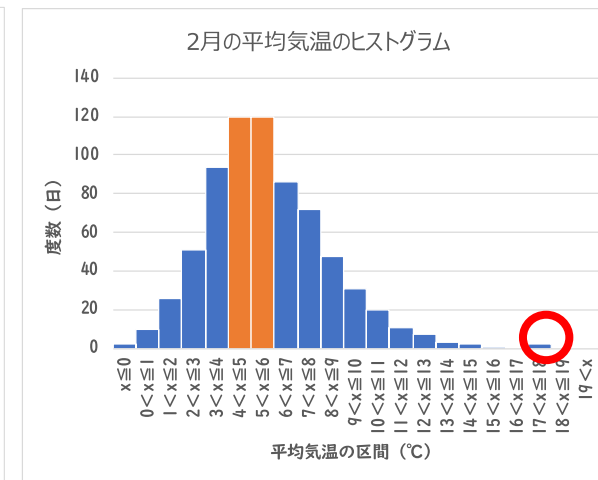
多峰



単峰



左右対称

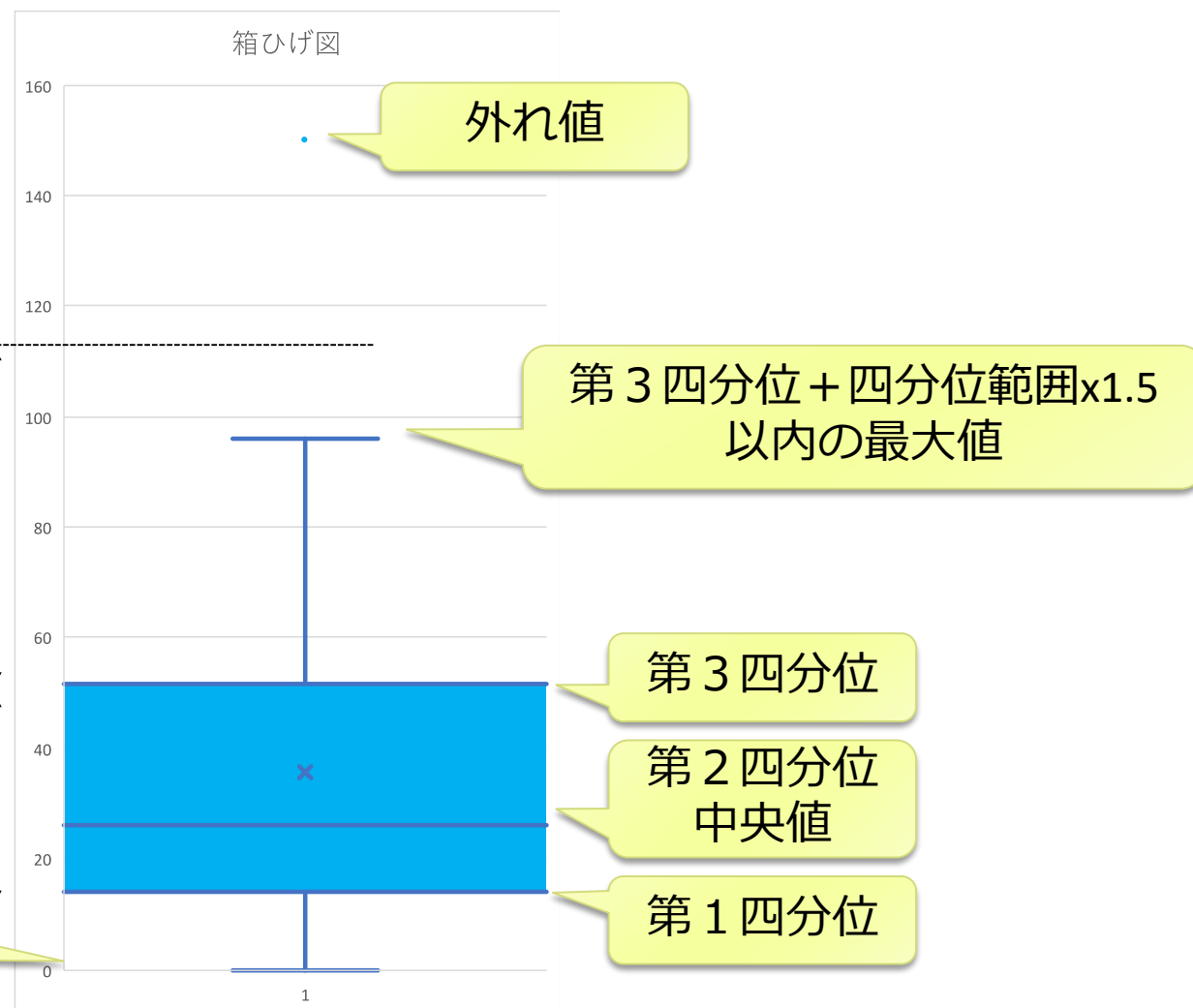
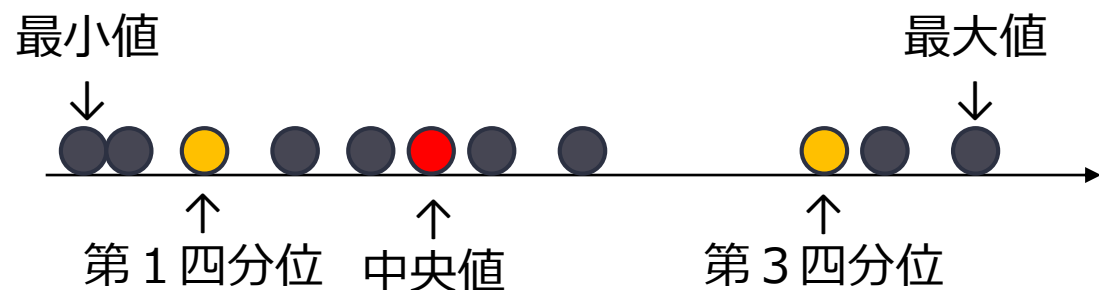


非対称・外れ値

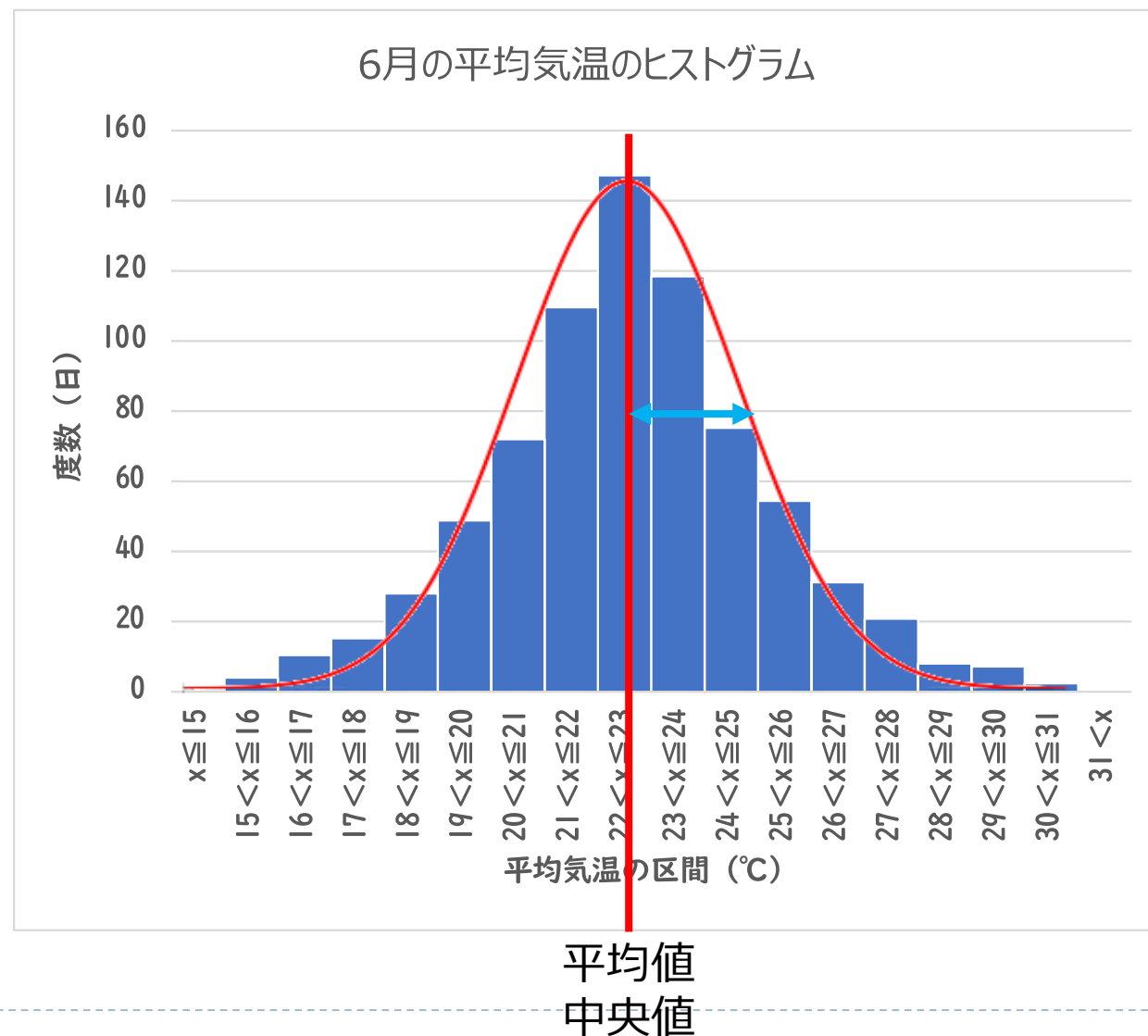
異なる要素を検討

基本統計量

歪んだ分布の中心



対称な分布の中心



平均値と中央値を求めるExcel関数

	A	B	C	D	E	F
1			東京			
2	年月日	月	平均気温(°C)			
3	1998/8/1	8	26.4	19.1	平均値	27.7351
4	1998/8/2	8	29.5	19.3	中央値 (メジアン)	28.2
5	1998/8/3	8	29.5	19.3		
6	1998/8/4	8	28.2	19.5		
7	1998/8/5	8	28.2	19.5		
8	1998/8/6	8	27.5	19.9		
9	1998/8/7	8	27.6	19.9		

=AVERAGE(C3:C777)

=MEDIAN(C3:C777)

データの取得（体力測定）



Q 検索語句を入力して下さい。 検索

ホーム

スポーツ庁について

お知らせ

政策

トップ > 刊行物 > 統計情報 > 全国体力・運動能力、運動習慣等調査 > 令和4年度全国体力・運動能力、運動習慣等調査結果

令和4年度全国体力・運動能力、運動習慣等調査結果

室伏長官のビデオメッセージ

調査結果を受けた室伏長官から皆さんへのビデオメッセージです。



スクロール

令和4年度全国体力・運動能力、運動習慣等調査 集計結果

小学校

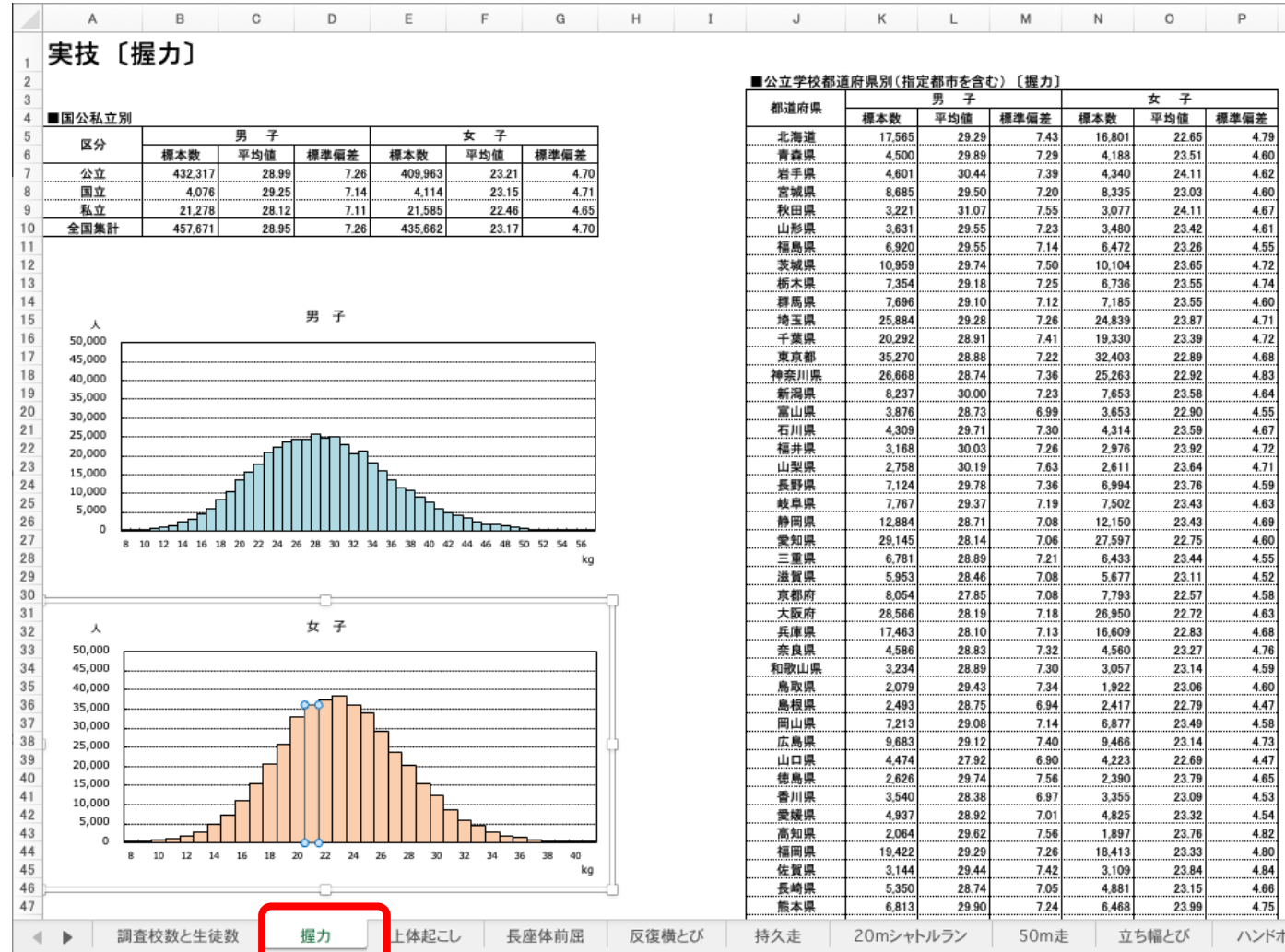
- 小学校 実技・体格(Excel:224KB) X
- 児童質問紙(Excel:258KB) X
- 学校質問紙(Excel:223KB) X

中学校

- 中学校 実技・体格(Excel:251KB) X
- 生徒質問紙(Excel:306KB) X
- 学校質問紙(Excel:231KB) X

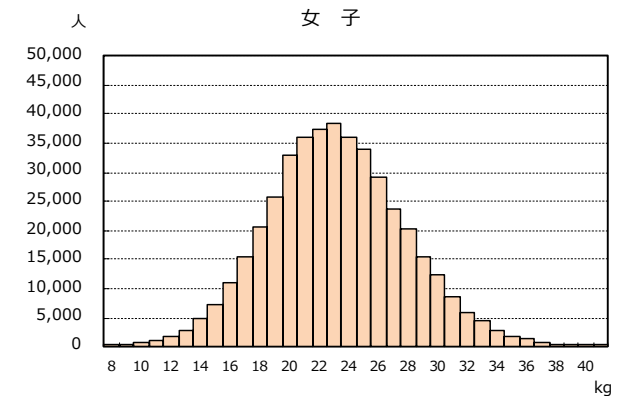
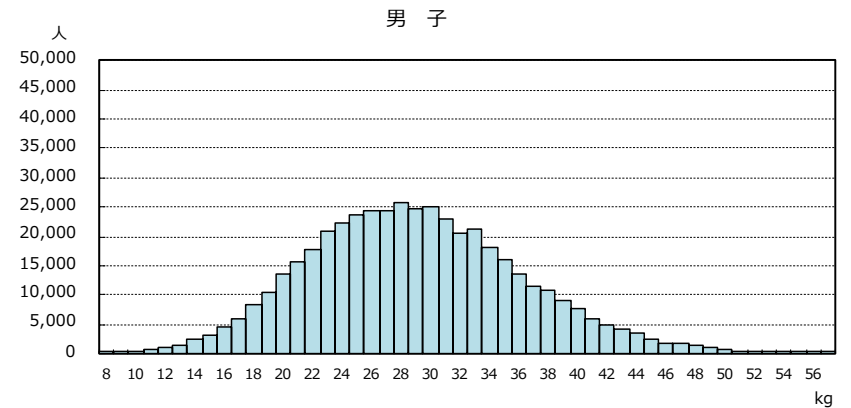
https://www.mext.go.jp/sports/b_menu/toukei/kodomo/zencyo/1411922_00004.html

体力測定データ



■公立学校都道府県別（指定都市を含む）〔握力〕

都道府県	男 子			女 子		
	標本数	平均値	標準偏差	標本数	平均値	標準偏差
北海道	17,565	29.29	7.43	16,801	22.65	4.79
青森県	4,500	29.89	7.29	4,188	23.51	4.60
岩手県	4,601	30.44	7.39	4,340	24.11	4.62
宮城県	8,685	29.50	7.20	8,335	23.03	4.60
秋田県	3,221	31.07	7.55	3,077	24.11	4.67
山形県	3,631	29.55	7.23	3,480	23.42	4.61
福島県	6,920	29.55	7.14	6,472	23.26	4.55
茨城県	10,959	29.74	7.50	10,104	23.65	4.72
栃木県	7,354	29.18	7.25	6,736	23.55	4.74
群馬県	7,696	29.10	7.12	7,185	23.55	4.80
埼玉県	25,884	29.28	7.26	24,839	23.87	4.71
千葉県	20,292	28.91	7.41	19,330	23.39	4.72
東京都	35,270	28.88	7.22	32,403	22.89	4.68
神奈川県	26,668	28.74	7.36	25,263	22.92	4.83
新潟県	8,237	30.00	7.23	7,653	23.58	4.64
富山県	3,876	28.73	6.99	3,653	22.90	4.55
石川県	4,309	29.71	7.30	4,314	23.59	4.67
福井県	3,168	30.03	7.26	2,976	23.92	4.72
山梨県	2,758	30.19	7.63	2,611	23.64	4.71
長野県	7,124	29.78	7.36	6,994	23.76	4.59
岐阜県	7,767	29.37	7.19	7,502	23.43	4.63
静岡県	12,884	28.71	7.08	12,150	23.43	4.69
愛知県	29,145	28.14	7.06	27,597	22.75	4.60
三重県	6,781	28.89	7.21	6,433	23.44	4.55
滋賀県	5,953	28.46	7.08	5,677	23.11	4.52
京都府	8,054	27.85	7.08	7,793	22.57	4.58
大阪府	28,566	28.19	7.18	26,950	22.72	4.63
兵庫県	17,463	28.10	7.13	16,609	22.83	4.68
奈良県	4,586	28.83	7.32	4,560	23.27	4.76
和歌山県	3,234	28.89	7.30	3,057	23.14	4.59
鳥取県	2,079	29.43	7.34	1,922	23.06	4.60
島根県	2,493	28.75	6.94	2,417	22.79	4.47
岡山県	7,213	29.08	7.14	6,877	23.49	4.58
広島県	9,683	29.12	7.40	9,466	23.14	4.73
山口県	4,474	27.92	6.90	4,223	22.69	4.47
徳島県	2,626	29.74	7.56	2,390	23.79	4.65
香川県	3,540	28.38	6.97	3,355	23.09	4.53
愛媛県	4,937	28.92	7.01	4,825	23.32	4.54
高知県	2,064	29.62	7.56	1,897	23.76	4.82
福岡県	19,422	29.29	7.26	18,413	23.33	4.80
佐賀県	3,144	29.44	7.42	3,109	23.84	4.84
長崎県	5,350	28.74	7.05	4,881	23.15	4.66
熊本県	6,813	29.90	7.24	6,468	23.99	4.75



福岡市	2
熊本市	2

■地域の規模別〔握力〕

区分	標本数
----	-----

個表 = ローデータ(raw)

度数から平均値（期待値）

	A	B	C	D
1	男子			
2	kg	人	累積	積
3	8	198	198	1584
4	9	228	426	2052
5	10	462	888	4620
6	11	659	1547	7249
7	12	1016	2563	12192
8	13	1508	4071	19604
9	14	2307	6378	32298
10	15	3251	9629	48765
11	16	4537	14166	72592
12	17	5084	19250	100000
53	合計	457671		13251209
54	中央値	228835.5	29	
55	平均値		28.95357	
56	最頻値		28	

=Kg × 人

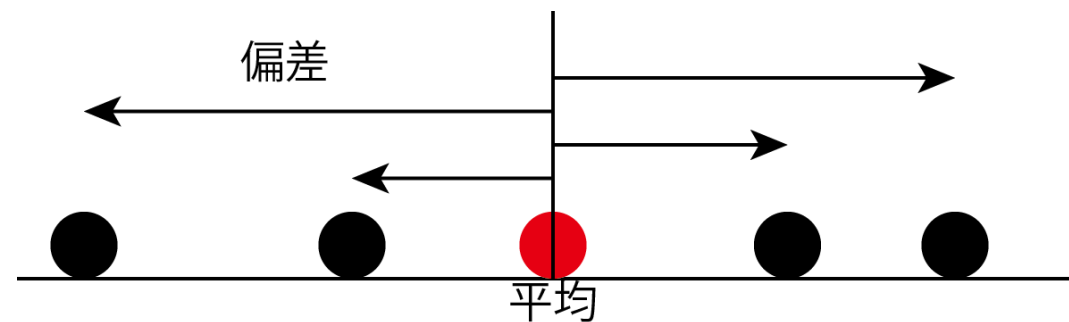
=上 × 左

G	H	I	J
女子			
kg	人	累積	積
8	265	265	2120
9	410	675	3690
10	672	1347	6720
11	1145	2492	12595
12	1757	4249	21084
13	2892	7141	37596
14	4730	11871	66220
15	7450	19321	111750
16	10960	30281	175360
17	15460	45741	262820
合計	435662		10094449
中央値	217831	23	
平均値		23.17037	
最頻値		23	

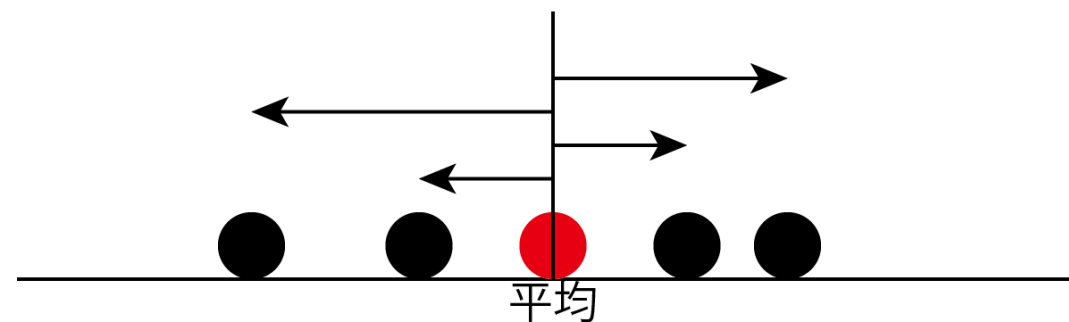
=Kg × 人

=上 × 左

偏差



(a) 散らばりが大きいデータ



(b) 散らばりが小さいデータ

平均と分散と標準偏差

平均

$$\bar{x} = \frac{(x_1 + x_2 + x_3 + \cdot \cdot \cdot + x_n)}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

n : データの個数

分散

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdot \cdot \cdot + (x_n - \bar{x})^2}{n} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

標準偏差

$$\sigma = \sqrt{s^2}$$

標準偏差を求めるExcel関数

	A	B	C	D	E	F	
2	年月日	月	平均気温(°C)				
3	1998/8/1	8	26.4	19.1	平均値	27.7351	
4	1998/8/2	8	29.5	19.3	中央値 (メジアン)	28.2	
5	1998/8/3	8	29.5	19.3	標準偏差	2.438743	
6	1998/8/4	8	28.2	19.5	標本標準偏差	2.440317	
7	1998/8/5	8	28.2	19.5			
8	1998/8/6	8	27.5	19.9			
9	1998/8/7	8	27.6	19.9			

=STDEV.P(C3:C777)

=STDEV.S(C3:C777)

不偏分散

標準偏差を求める

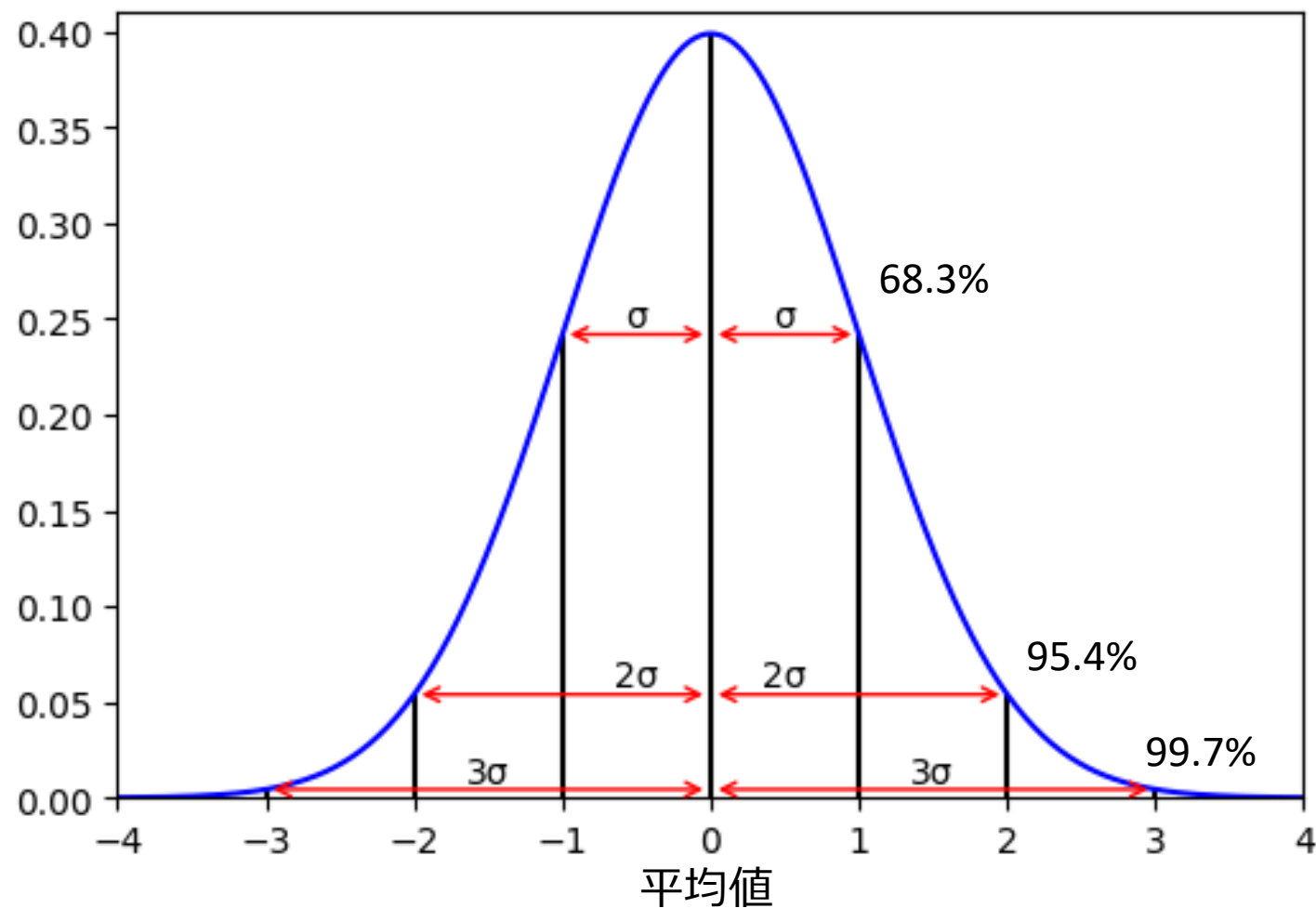
	A	B	C	D	E	F
1	男子					
2	kg	人	累積	積	偏差^2	偏差^2積
3	8	198	198	1584	439.05207	86932.309
4	9	228	426	2052	398.14493	90777.043
5	=(Kg-\$平均)^2				23779	165967.86
6					33065	212415.9
7	12	1016	=偏差^2 * 人			
8	13	1508				
9	14	2307	6378	32298	223.60923	515866.5
10	15	3251	9629	48765	194.7021	632976.51

53	合計	4509	=SUM(F4:F53)		24103151
54	中央値	228	=合計/人数		
55	平均値		分散		52.664799
56	最頻値	28	標準偏差		7.2570517
			=SQRT(分散)		

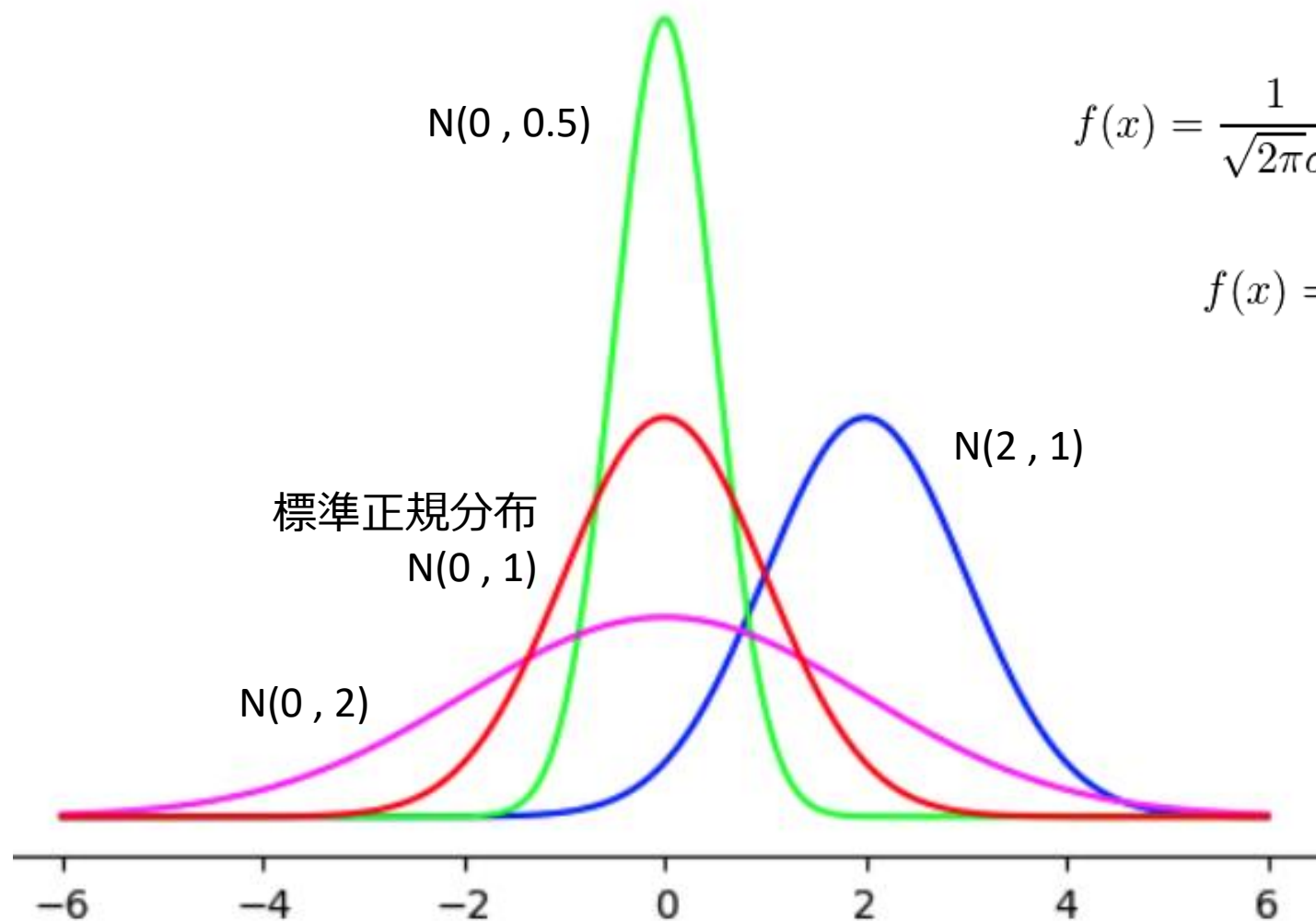
	G	H	I	J	K	L
	女子					
	kg	人	累積	積	偏差^2	偏差^2積
	8	265	265	2120	230.1401	60987.12
	9	410	675	3690	200.7993	82327.73
	10	672	1347	6720	173.4586	116564.2
	=(Kg-\$平均)^2				1179	169595
	12	1757	4249	21084	124.7771	219233.4
	13	2892	=偏差^2 * 人			
	14	4730	11871	66220	84.09565	397772.4
	15	7450	19321	111750	66.75492	497324.1
	16	10960	30281	175360	51.41418	563499.4

合計	435662		10094449		9607634
中央値	217831	23			
平均値		23.17037		分散	22.052953
最頻値		23		標準偏差	4.6960572

正規分布と標準偏差



正規分布



$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (-\infty < x < \infty)$$

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \quad (-\infty < x < \infty)$$

要約

正規分布

項目	値
データ数	457,671
平均	28.95
標準偏差	7.26
平均 - 標準偏差	21.70
平均 + 標準偏差	36.21
その間のデータ数	303,151
割合	66.2%

18	21-22kg	22	17834	86286	392348	862311.8071
19	22-23kg	23	20960	107246	482080	742926.9279
20	23-24kg	24	22267	129513	534408	546384.2722
21	24-25kg	25	23650	153163	591250	369666.2903
22	25-26kg	26	24359	177522	633334	212497.4761
23	26-27kg	27	24303	201825	656181	92750.76831
24	27-28kg	28	25743	227568	720804	23407.96437
25	28-29kg	29	24805	252373	719345	53.47494063
26	29-30kg	30	25108	277481	753240	27493.69384
27	30-31kg	31	20492	300431	711450	96111.8172
28	31-32kg	32	21124	320923	655744	190180.9284
29	32-33kg	33	17960	342047	697092	345875.9618
30	33-34kg	34	15987	375994	559545	584473.9225
31	34-35kg	35	13443	389437	483948	667474.3374
32	35-36kg	36	11362	400799	420394	735633.2302
33	36-37kg	37				

zスコアと偏差値

$$z_i = \frac{x_i - \mu}{\sigma}$$

μ : 平均
 σ : 標準偏差

$$\text{偏差値} = z_i \times 10 + 50$$

男子の握力25kg

握力	zスコア	偏差値
25	-0.54	44.55

女子の握力25kg

握力	zスコア	偏差値
25	0.39	53.90

変動係数(CV : coefficient of variation)

$$CV = \frac{\text{標準偏差}}{\text{平均値}}$$

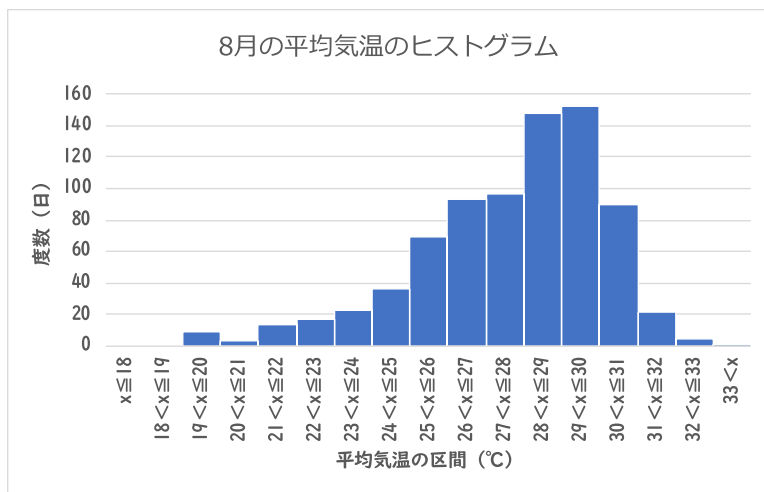
握力	男子	女子
平均	28.95	23.17
標準偏差	7.26	4.70
CV	0.25	0.20

長座体前屈	男子	女子
平均	43.76	46.05
標準偏差	11.22	10.54
CV	0.26	0.23

歪度

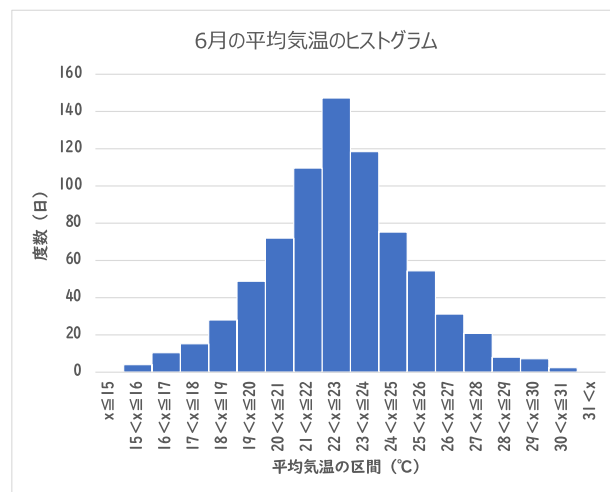
▶ 分布の歪みの方向性

$$\frac{1}{n} \sum_{i=1}^n z_i^3$$



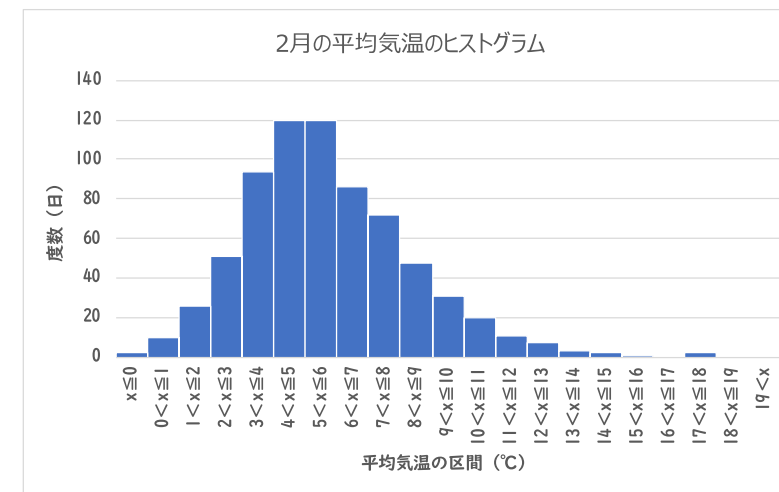
歪度

-0.94



歪度

0.10



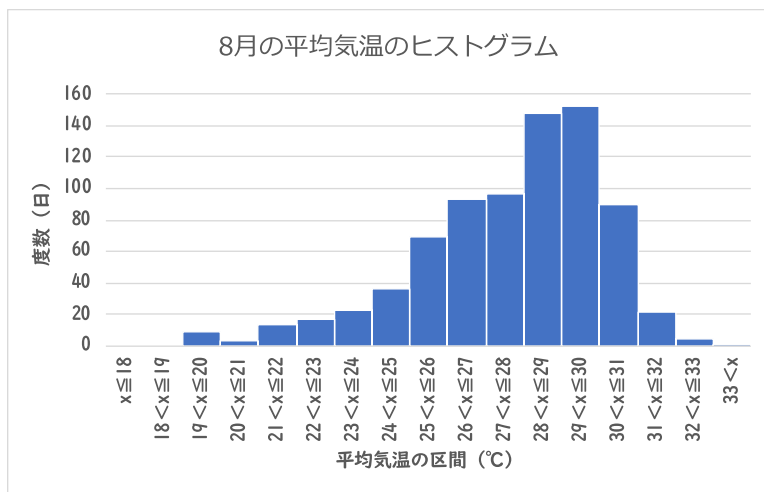
歪度

0.76

尖度

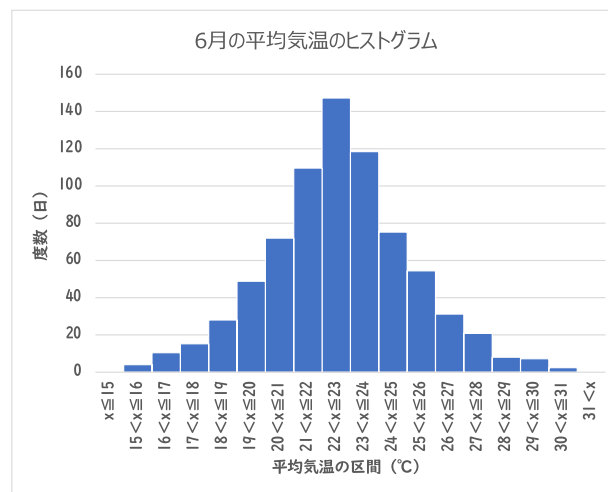
▶ 山の尖り具合と裾の広がり

$$\frac{1}{n} \sum_{i=1}^n z_i^4$$



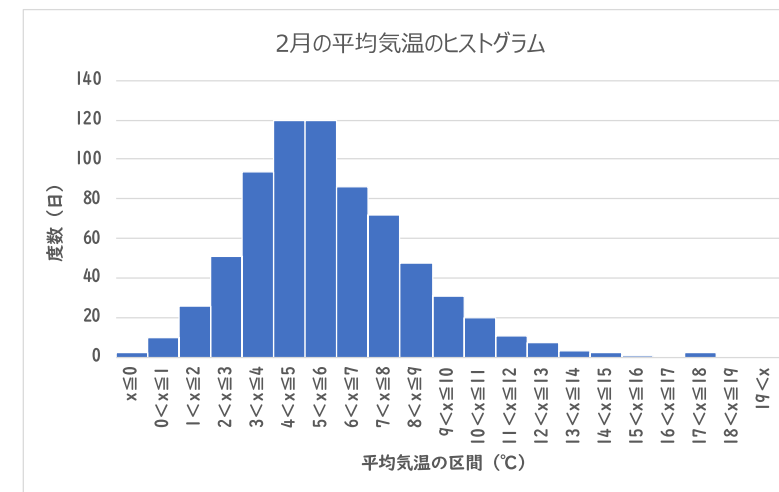
尖度

0.90



尖度

0.34

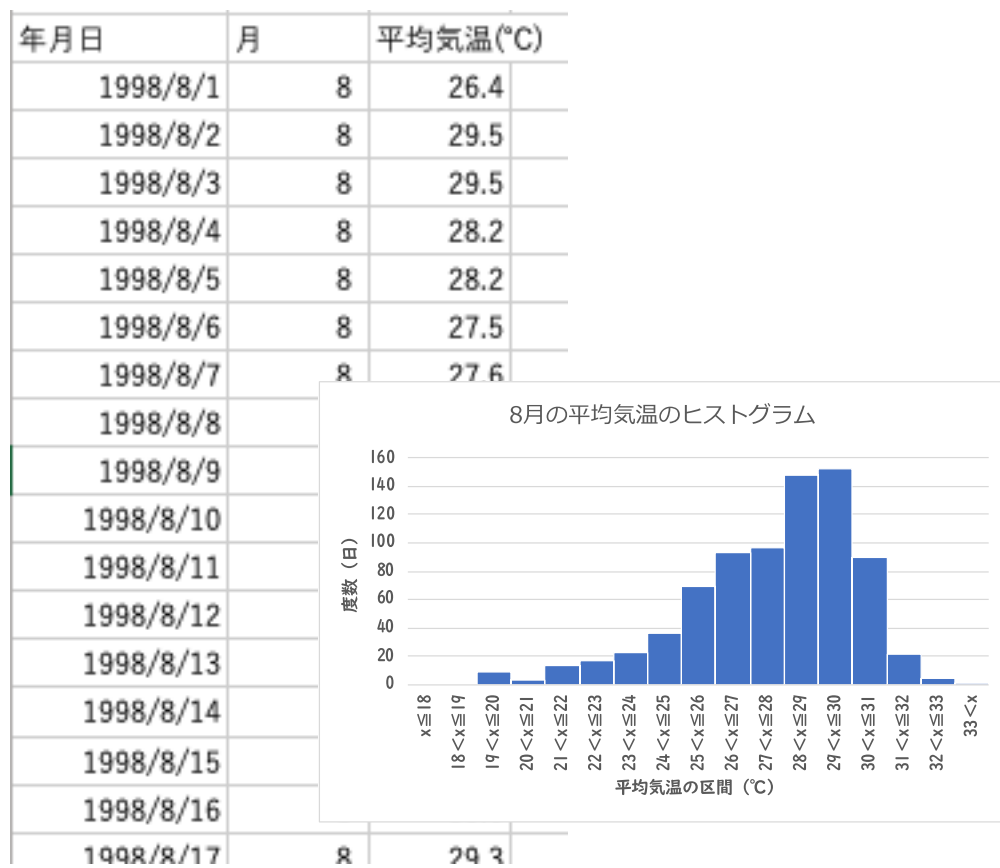


尖度

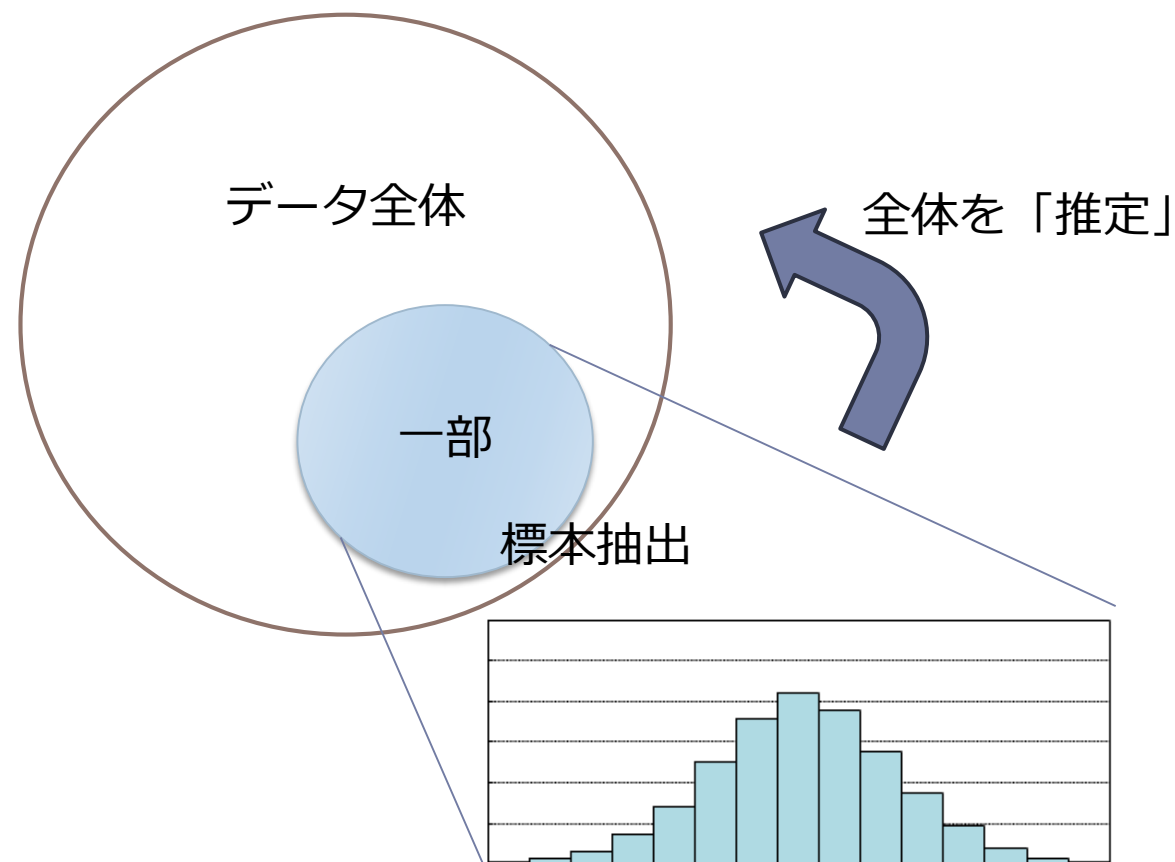
1.20

記述統計と推測統計

記述統計

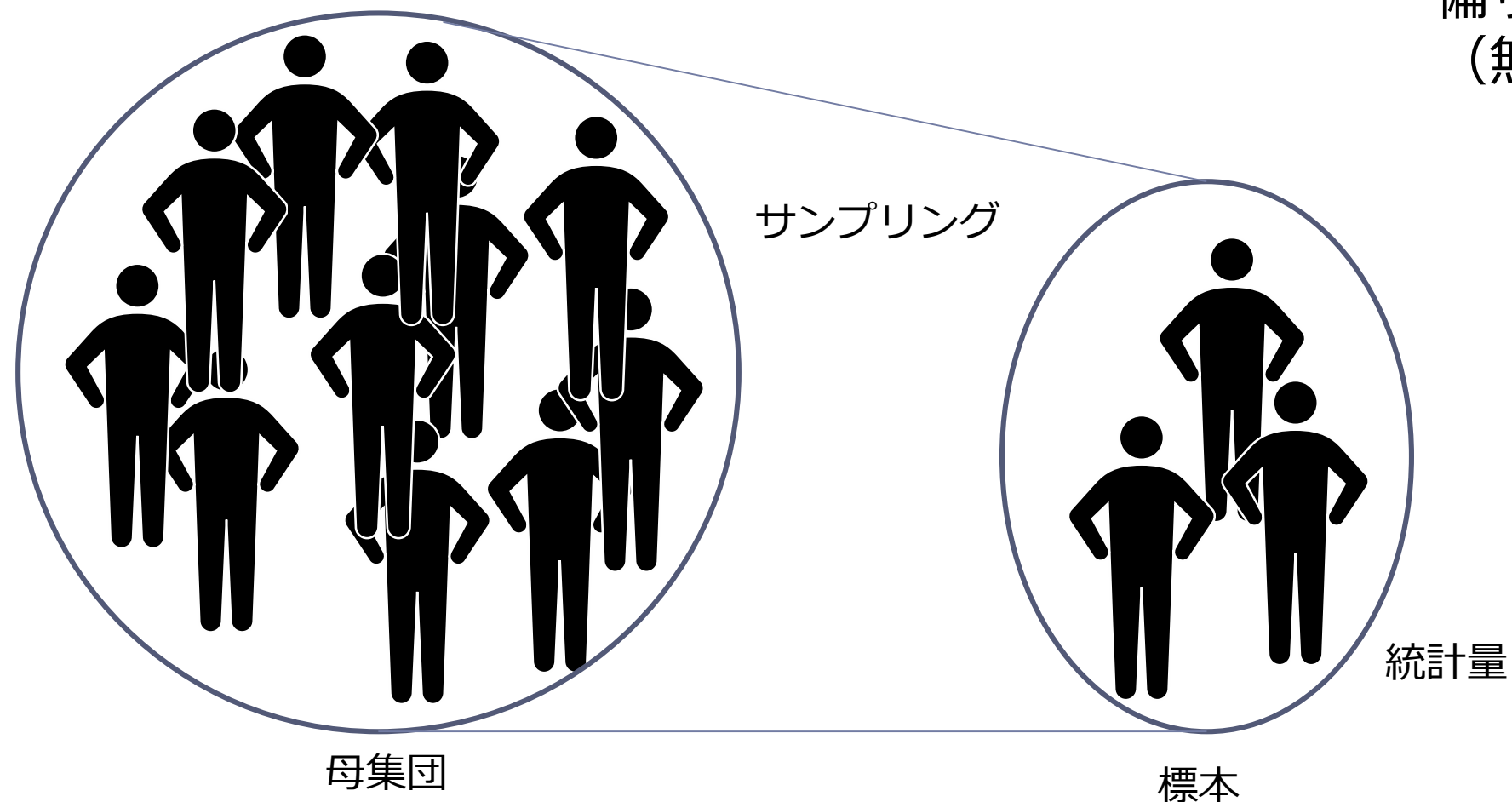


推測統計



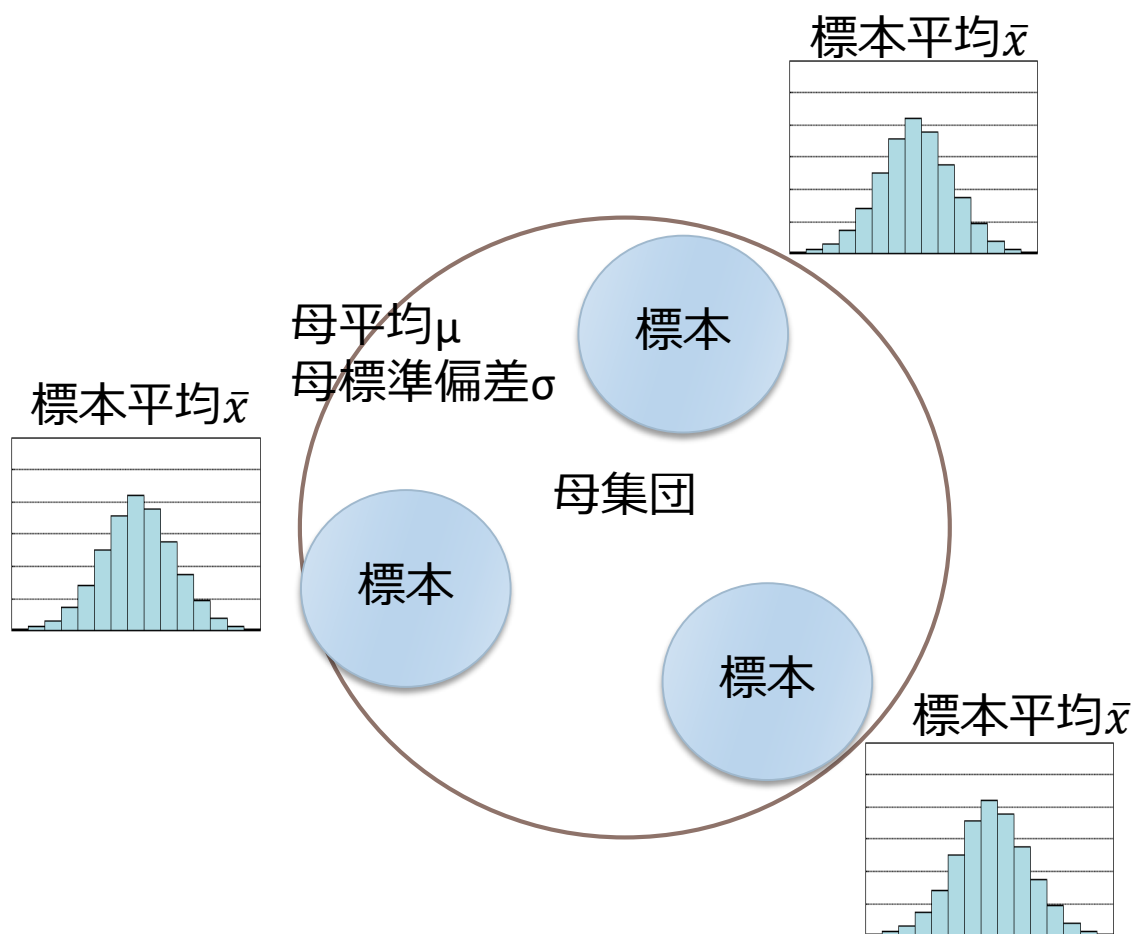
推測統計

標本を調べて、母集団を推測する



標本分布

母集団から複数の標本

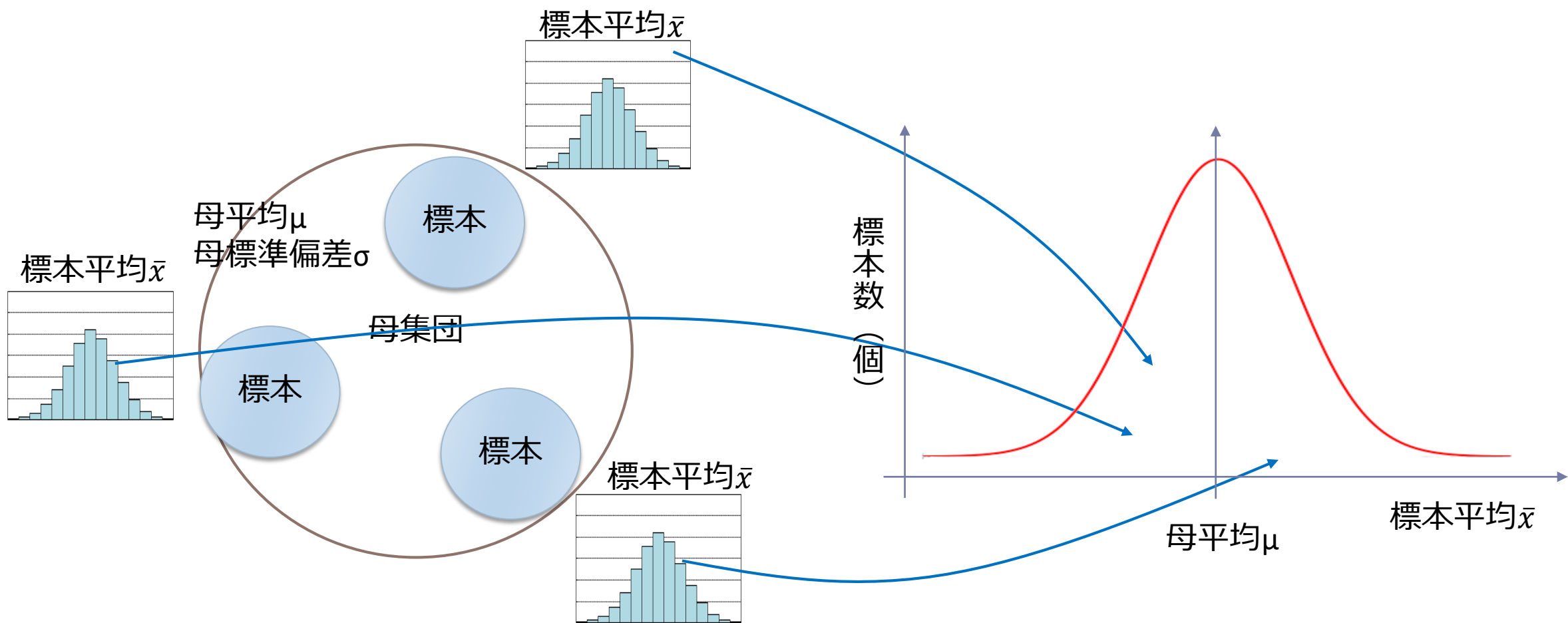


標本の分布

標本分布

母集団から複数の標本

標本の分布



標本（2月の平均気温）

S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
番号	乱数	乱数の固定値	該当データ	乱数	該当データ	乱数	該当データ	乱数	該当データ	乱数	該当データ
1	246	606	12	277	3.7	589	11.3	76	7.5	482	5
2	202	609	10.3	284	6.2	570	5.7	529	6.1	252	9.4
3	64	477	6.6	118	5.2	165	12	280	6.4	573	0.8
4	46	542	3.3	137	6	367	6	573	0.8	345	8.2
5	271	319	11.8	48	5.2	251	6.1	356	9.1	692	3.5
6	217	81	5.9	600	6.2	658	6.2	178	7.5	271	5.3
7	208	413	5.2	217	6.4	387	8.1	37	6.1	178	7.5
8	208	605	11.7	158	8.8	139	8.1	256	1.5	253	10

50個ランダムに取り出す

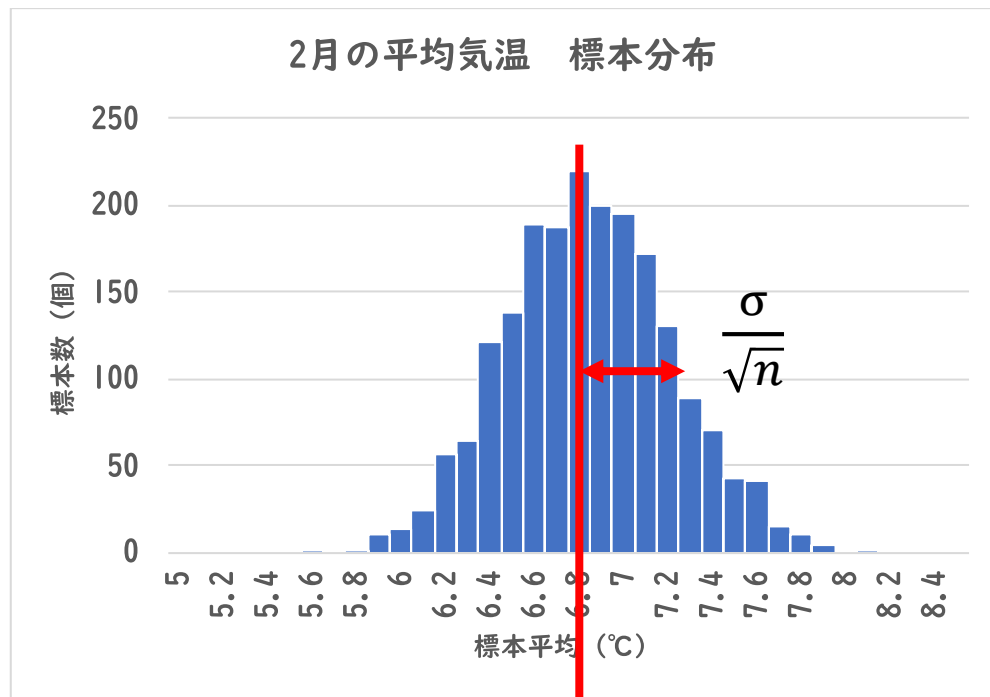
	平均	6.44	平均	6.754	平均	7.084	平均	6.608	平均	6.726
	標準偏差	2.17	標準偏差	2.36	標準偏差	2.17	標準偏差	2.16	標準偏差	2.65
	平均の平均	6.7224								

平均の平均

各標本の平均と標準偏差

標本分布

▶ 標本平均のヒストグラム



$$\frac{\sigma}{\sqrt{n}}$$

標準誤差(Standard Error : SE)

▶ 標本平均の特徴

1. 標本平均の平均 $\bar{x} =$ 母集団の平均 μ
2. 標本分布の分散 $\frac{\sigma^2}{n}$ (n : 標本サイズ)
3. 母集団が正規分布 \rightarrow 標本分布も正規分布
4. n が大きければ標本分布も正規分布

項目	標本分布	母集団
平均	6.80	6.80
中央値	6.80	6.40
最頻値	6.80	6.01
分散	0.138	6.94

$$\frac{6.94}{50} = 0.139$$

点推定

- ▶ 1つの標本の平均を母平均とみなす

標本分散

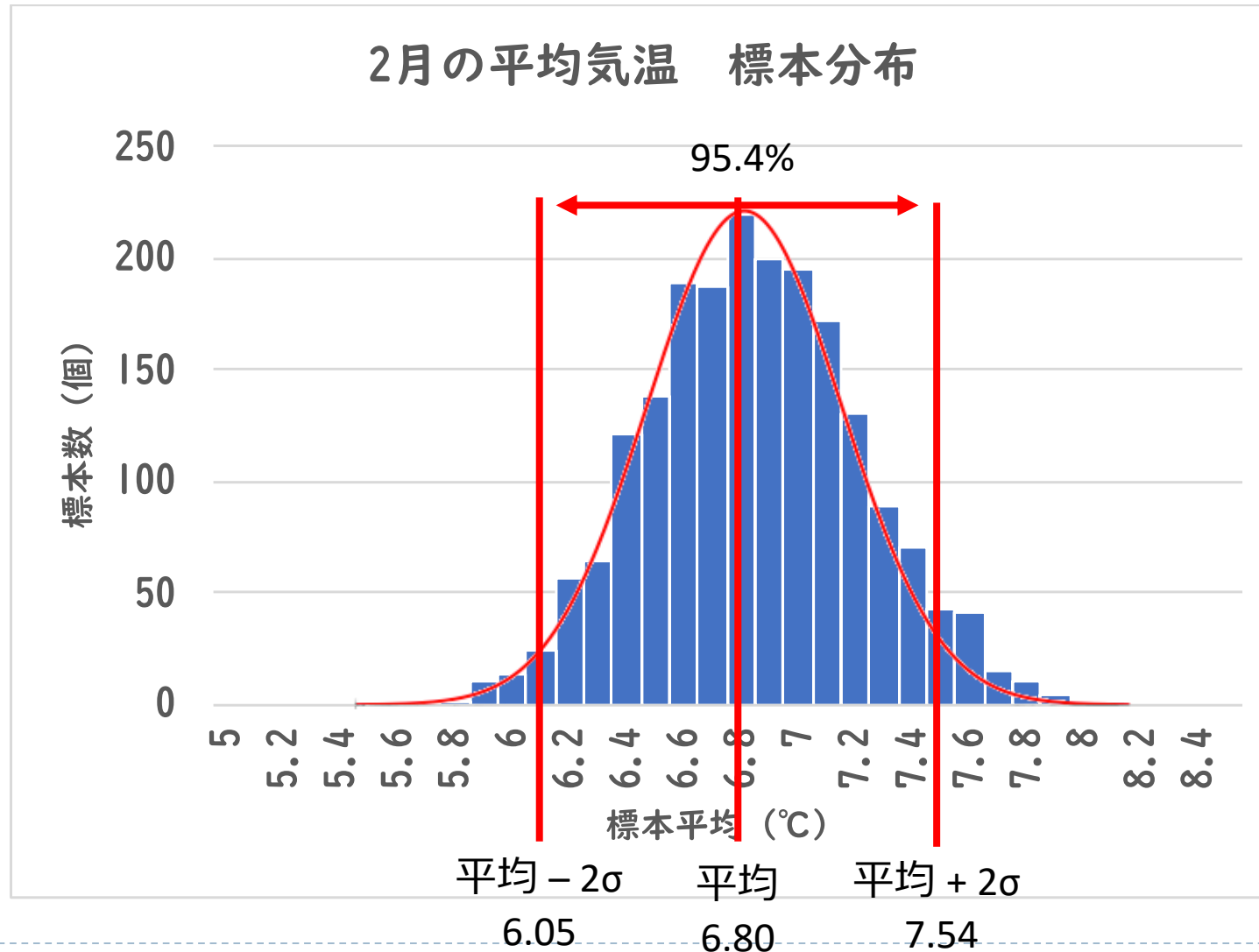
$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

不偏分散

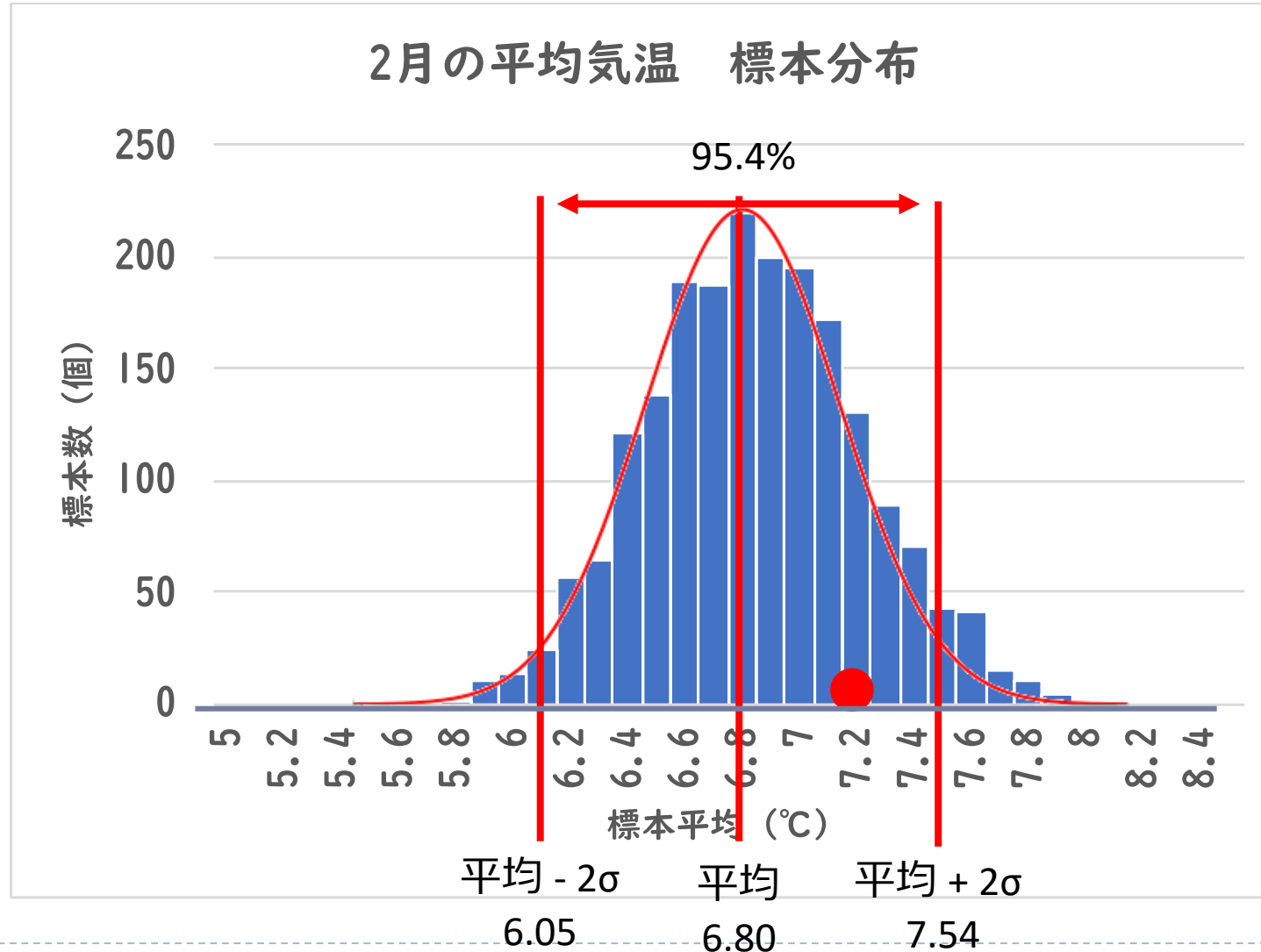
$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

区間推定



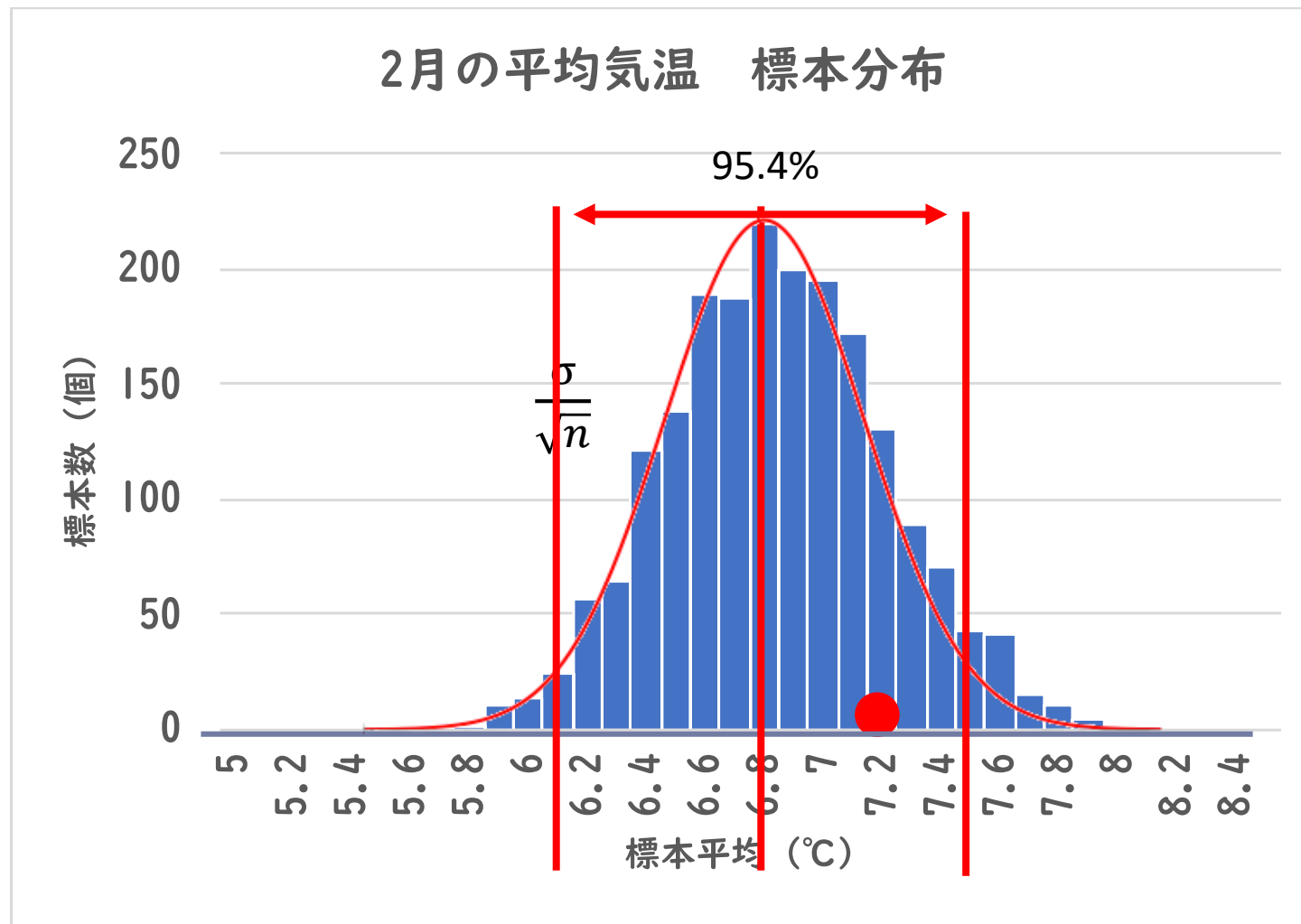
標本分布の平均	6.80
標本分布の標準偏差 σ	0.37

区間推定



標本分布の平均	6.80
標本分布の標準偏差 σ	0.37

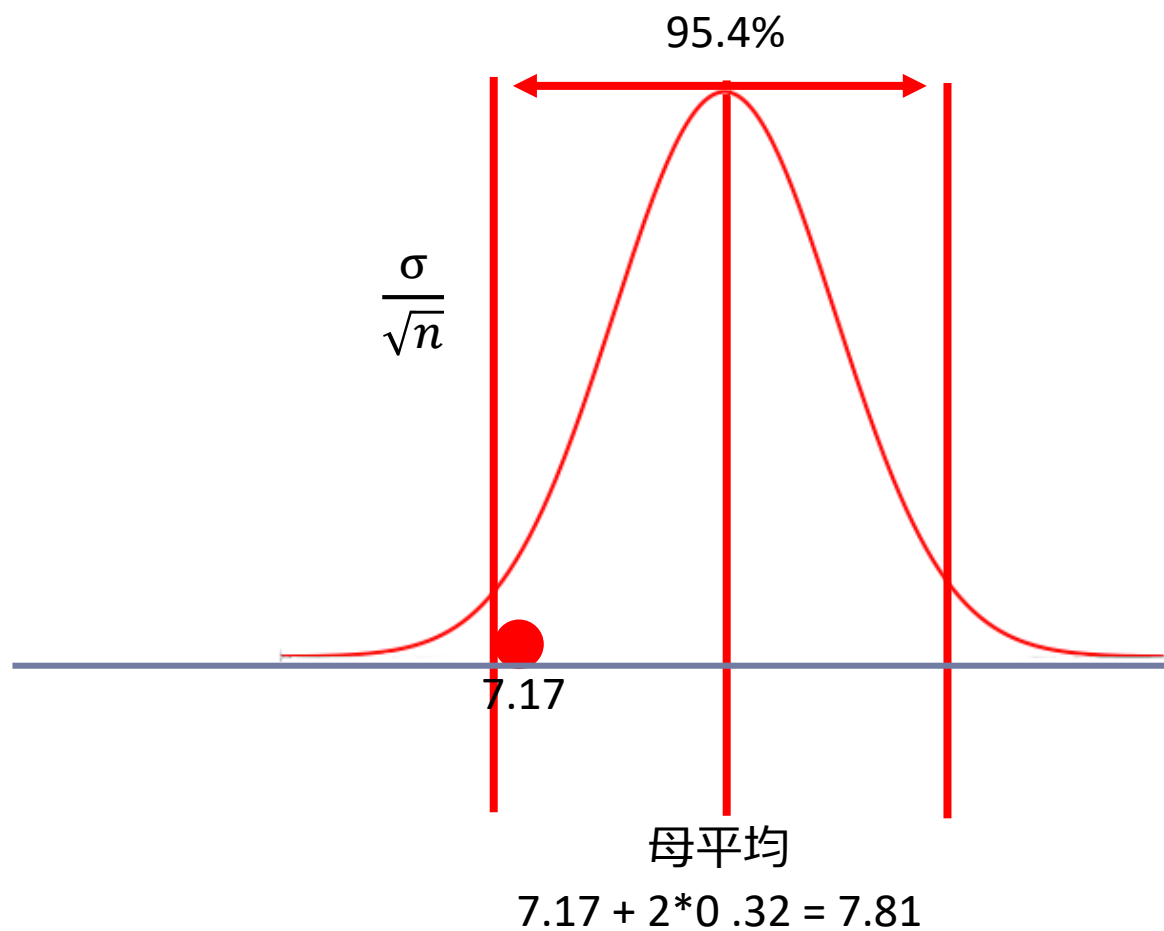
区間推定



$$S.E. = \frac{\sigma}{\sqrt{n}}$$

標本分布の平均	6.80
標本分布の標準偏差	0.37
標本の不偏分散	5.16
標準誤差	0.32

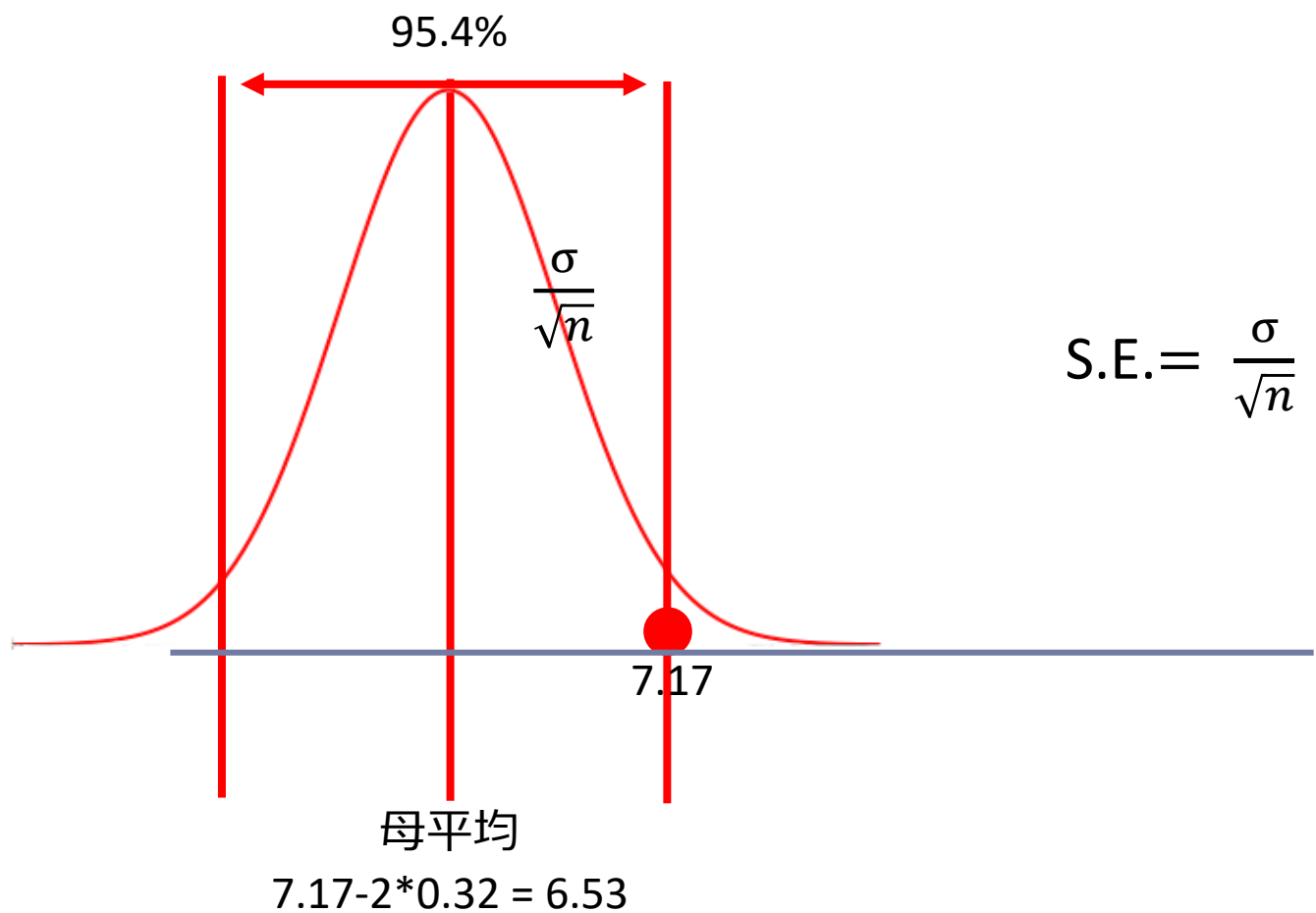
区間推定



$$S.E. = \frac{\sigma}{\sqrt{n}}$$

標本分布の平均	6.80
標本分布の標準偏差	0.37
標本の不偏分散	5.18
標準誤差	0.32

区間推定

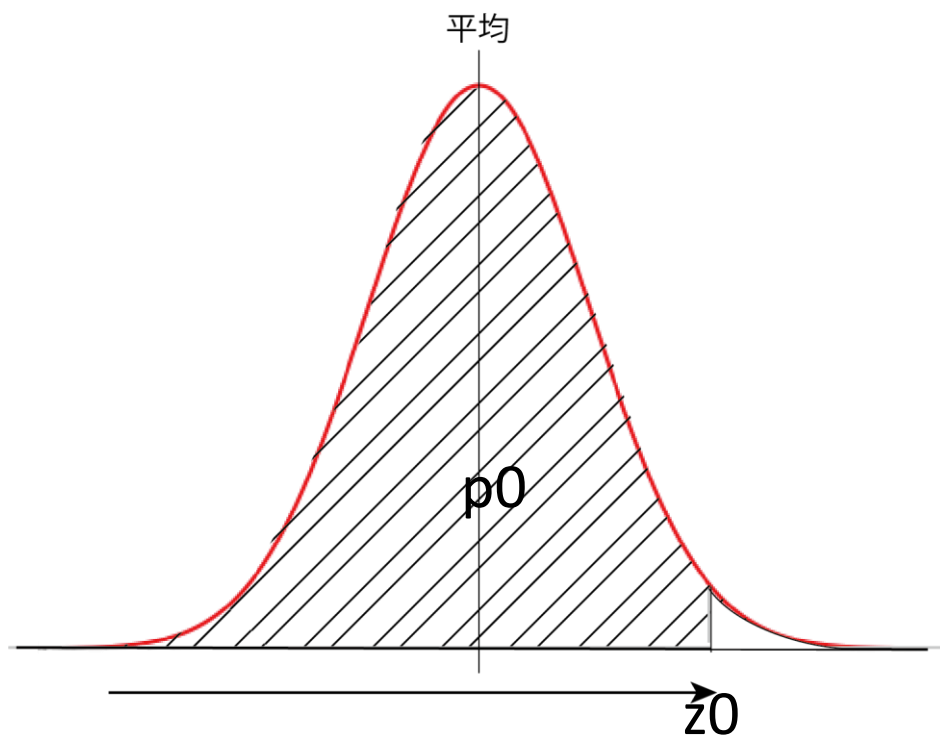


標本分布の平均	6.80
標本分布の標準偏差	0.37
標本の不偏分散	5.18
標準誤差	0.32

標準正規分布

▶ • $\text{NORM.S.DIST}(z_0) \rightarrow p_0$

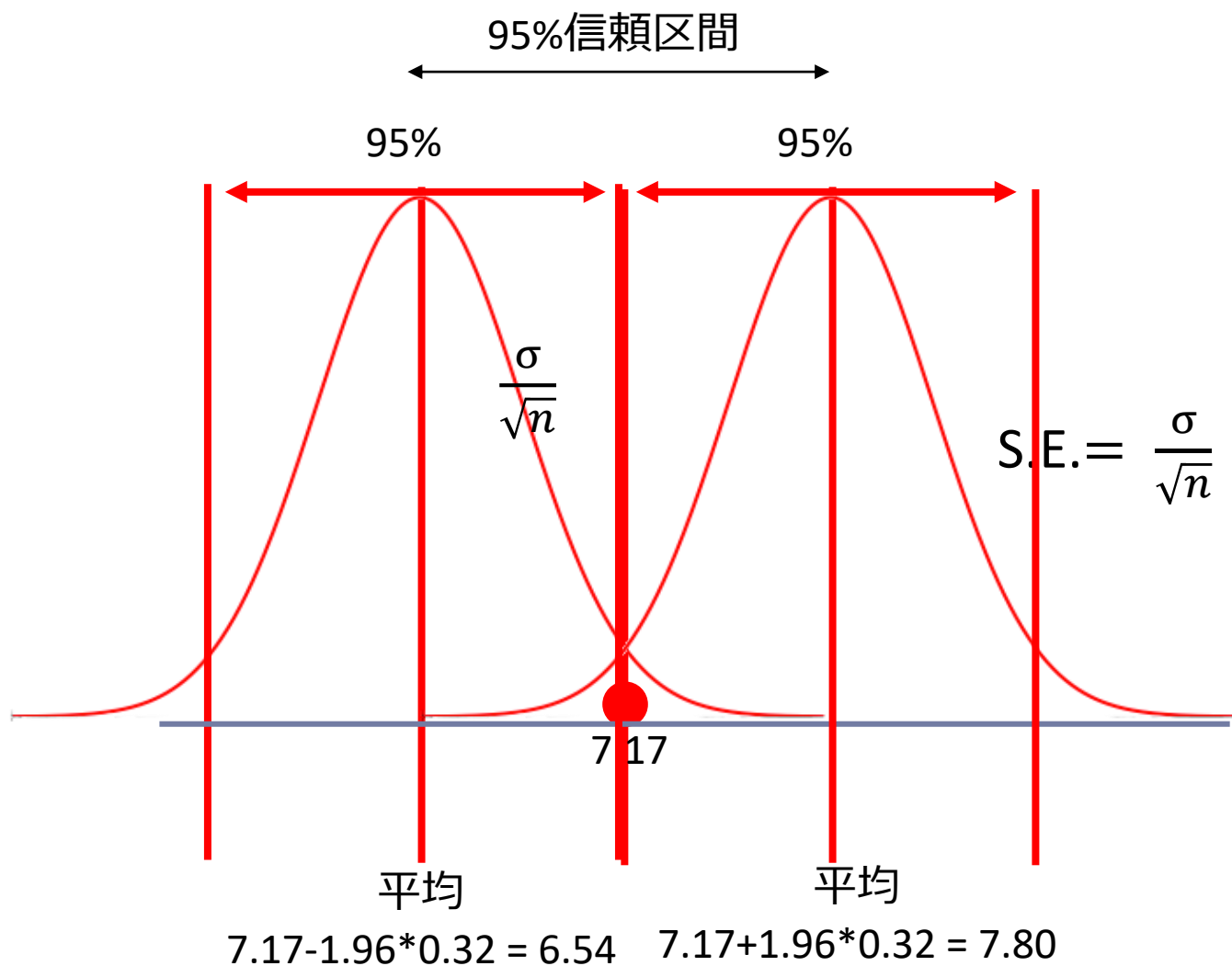
▶ • $\text{NORM.S.INV}(p_0) \rightarrow z_0$



$=\text{NORM.S.DIST}(2)$	0.977
$=0.977-0.5$	0.478
$=0.478 * 2$	0.954

$=0.95 / 2$	0.475
$=0.475+0.5$	0.975
$=\text{NORM.S.INV}(0.975)$	1.96

区間推定



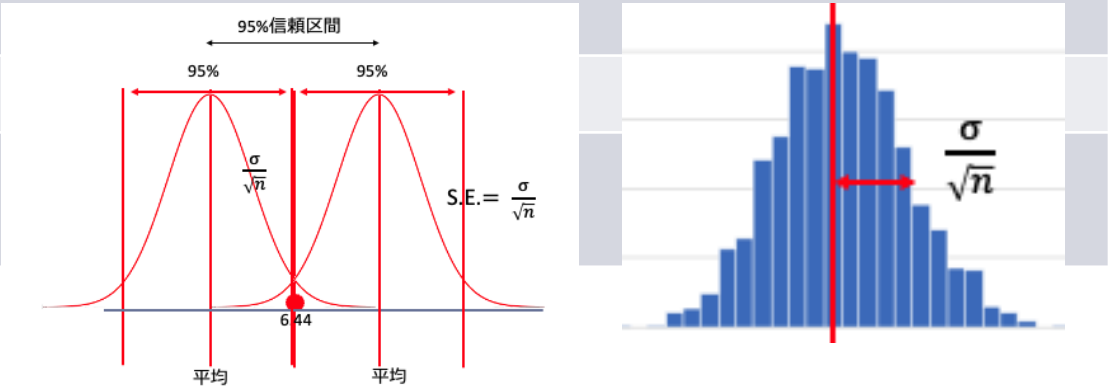
標本分布の平均	6.80
標本分布の標準偏差	0.37
標本の不偏分散	5.18
標準誤差	0.32

まとめ

項目	内容		
記述統計			
期待値	$\frac{1}{n} \sum_{i=1}^n x_i$	=AVERAGE(全データ)	
分散	$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$	=VAR.P(全データ)	
標準偏差	$\sqrt{s^2}$	=STDEV.P(全データ)	
歪度	$\frac{1}{n} \sum_{i=1}^n z_i^3$	尖度	$\frac{1}{n} \sum_{i=1}^n z_i^4$
推測統計			
標本分布	標本をたくさん作って分布にする		
点推定	母平均≒標本平均		
区間推定	母平均が95%の確率でこの範囲にある区間		

$$\bar{x} - 1.96S.E \leq \mu \leq \bar{x} + 1.96S.E$$

$$S.E = \frac{\sigma}{\sqrt{n}}$$



参考文献

- ▶ 「身近な統計」 石崎克也・渡辺美智子 放送大学教育振興会
- ▶ 「誰でも使える統計オープンデータ」 オフィシャル スタディノート
総務省統計局