５．統計・分析

数学Ⅰ第４章（数式上の理解）と連携し、下記に述べる「実社会における実践活用」の基礎を学ぶ

データマイニング：ＤＢから有用な情報を抽出する技術体系
データから，高頻度で発生する特徴的なパターンを見つける。

相関ルール抽出：ＤＢから、頻繁に同時に生起する事象同士を相関の強い事象の関係、すなわち相関ルールとして抽出する技術。ＰＯＳ、ＥＣ取引ログに含まれる購買履歴を利用した解析
例：スーパーでオムツを買った人のうちビールを買う人が多い → 両者を同じ場所に置く。
例：本Ａを買う人は、後に本Ｂを買うことが多い → 本Ａの購入者に本Ｂを薦めるメールを送る。

クラス分類：与えられたデータに対応するカテゴリを予測する問題
代表的な手法：単純ベイズ分類器，決定木，サポートベクターマシン
例：薬品の化合物のデータから，その化合物に薬効がある・ないといったカテゴリを予測

回帰分析：与えられたデータに対応する実数値を予測する問題
代表的な手法：線形回帰、ロジスティック回帰、サポートベクトル回帰
例：曜日、降水確率、今日の売上げなどのデータを元に、明日の売上げという実数値データを予測

クラスタリング：データの集合をクラスタと呼ぶグループに分ける。クラスタとは、同じクラスタのデータならば互いに似ていて、違うクラスタならば似ていないようなデータの集まり。
例：ウェブ閲覧パターンのデータから、類似したものをまとめることで、閲覧の傾向が同じ利用者のグループを発見する。

5-1　統計基礎実習　代表値・標準偏差・偏差値・ヒストグラム・相関・四分位数（箱ひげ図）

課題　母集団「A」「B」に対して合計,個数,最小値,最大値,中央値,平均値,最頻値、頻度を求めた後に、偏差,(偏差)^2,分散,標準偏差,偏差値を求め、さらにヒストグラム（度数分布をグラフで表したもの）を描き母集団の傾向を比較する

集団A							
母集団A	ポイント	偏差	(偏差)^2	(偏差)^2 平均値	(偏差)^2 平均値の平方根	偏差値	順位
A1	42						
A2	25						
A3	36						
A4	38						
A5	55						
A6	14						
A7	65						
A8	67						
A9	78						
A10	63						
A11	39						
A12	59						
A13	57						
A14	86						
A15	53						
A16	75						
A17	48						
A18	45						
A19	86						
A20	29						

偏差＝(データ)－(平均値)
偏差合計＝0
分散＝(偏差)^2 の平均値
標準偏差＝分散の平方根

合計	
個数	
最小値	
最大値	
中央値	
平均値	
最頻値	
分散	
標準偏差	
区間・階級	頻度
9.9	
19.9	
29.9	
39.9	
49.9	
59.9	
69.9	
79.9	
89.9	
99.9	
100	
最小値	
第1四分位点	
中央値	
第3四分位点	
最大値	

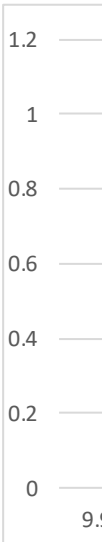
表計算ソフトの統計に関する主な関数

関数名	関数	目的
個数	COUNT	データの個数を求める
最小値	MIN	数値またはデータの最小値を求める
最大値	MAX	数値またはデータの最大値を求める
中央値	MEDIAN	数値の中央値を求める
平均値	AVERAGE	数値またはデータの平均値を求める
最頻値	MODE	数値の最頻値を求める
分散	VAR.P	数値をもとに分散を求める
標準偏差	ATDEV.P	数値をもとに標準偏差を求める
順位	RANK	順位を求める（同じ値のときは最上位の順位を返す）
度数分布	FREQENCY	区間に含まれる値の個数を求める
検索個数	COUNTIF	条件に一致するデータの個数を求める
平方根	SQRT	数値の正の平方根を求める
四分位点	QUARTILE	=（データ範囲,戻り値 0～4)
偏差値	右式	$\frac{\{(データ)－(平均値)\}\times 10}{標準偏差}+50$

↑注　Excel の泣き所：頻度算出のための区間設定をm以上ｎ未満にできない（m以上ｎ以下になる）
COUNTIF やマクロを使うと可能であるが煩雑になるため、ここではその処理を省く

集団B							
母集団B	ポイント	偏差	$(偏差)^2$	$(偏差)^2$ 平均値	$(偏差)^2$ 平均値の平方	偏差値	順位
B1	45						
B2	54						
B3	72						
B4	33						
B5	64						
B6	36						
B7	42						
B8	56						
B9	51						
B10	65						
B11	66						
B12	55						
B13	49						
B14	28						
B15	61						
B16	42						
B17	52						
B18	54						
B19	56						
B20	79						

合計	
個数	
最小値	
最大値	
中央値	
平均値	
最頻値	
分散	
標準偏差	
区間・階級	頻度
9.9	
19.9	
29.9	
39.9	
49.9	
59.9	
69.9	
79.9	
89.9	
99.9	
100.9	
最小値	
第1四分位点	
中央値	
第3四分位点	
最大値	



ポイント 65 の集団Ａにおける偏差値…＿＿＿＿＿＿＿＿

ポイント 65 の集団Ｂにおける偏差値…＿＿＿＿＿＿＿＿

平均値・中央値が同値の２つの集団においても＿＿＿＿＿＿＿＿が異なると偏差値が変化する。

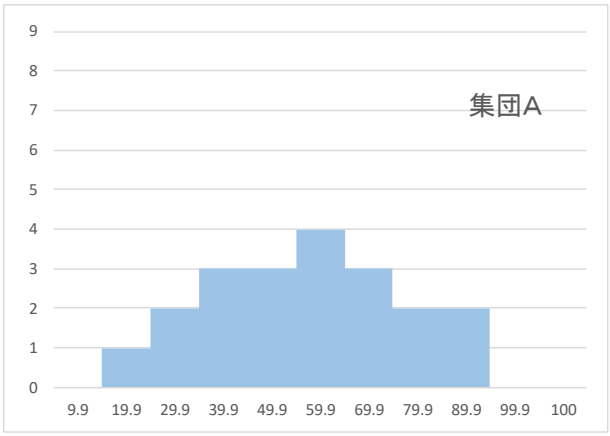
関数処理結果

集団A							
母集団A	ポイント	偏差	(偏差)^2	(偏差)^2 平均値	(偏差)^2 平均値の平方	偏差値	順位
A1	42	−11.0	121.0			44.4	14
A2	25	−28.0	784.0			35.6	19
A3	36	−17.0	289.0			41.3	17
A4	38	−15.0	225.0			42.3	16
A5	55	2.0	4.0			51.0	10
A6	14	−39.0	1521.0			30.0	20
A7	65	12.0	144.0			56.2	6
A8	67	14.0	196.0			57.2	5
A9	78	25.0	625.0			62.8	3
A10	63	10.0	100.0			55.1	7
A11	39	−14.0	196.0			42.8	15
A12	59	6.0	36.0			53.1	8
A13	57	4.0	16.0			52.1	9
A14	86	33.0	1089.0			66.9	1
A15	53	0.0	0.0			50.0	11
A16	75	22.0	484.0			61.3	4
A17	48	−5.0	25.0			47.4	12
A18	45	−8.0	64.0			45.9	13
A19	86	33.0	1089.0			66.9	1
A20	29	−24.0	576.0	379.2	19.5	37.7	18
合計	1060	0.0	7584.0			50.0	
個数	20		379.2				
最小値	14.0						
最大値	86.0						
中央値	54.0						
平均値	53.0		379.2		19.5		
最頻値	86.0						
分散	379.2						
標準偏差	19.5						
区間・階級	頻度						
9.9	0						
19.9	1						
29.9	2						
39.9	3						
49.9	3						
59.9	4						
69.9	3						
79.9	2						
89.9	2						
99.9	0						
100.9	0						
最小値	14						
第1四分位点	38.75						
中央値	54						
第3四分位点	65.5						
最大値	86						

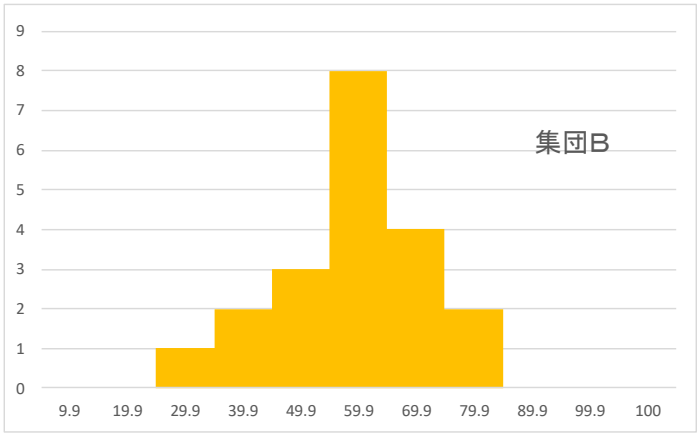
集団B							
母集団B	ポイント	偏差	(偏差)^2	(偏差)^2 平均値	(偏差)^2 平均値の平方	偏差値	順位
B1	45	−8.0	64.0			43.7	15
B2	54	1.0	1.0			50.8	10
B3	72	19.0	361.0			65.0	2
B4	33	−20.0	400.0			34.2	19
B5	64	11.0	121.0			58.7	5
B6	36	−17.0	289.0			36.6	18
B7	42	−11.0	121.0			41.3	16
B8	56	3.0	9.0			52.4	7
B9	51	−2.0	4.0			48.4	13
B10	65	12.0	144.0			59.5	4
B11	66	13.0	169.0			60.3	3
B12	55	2.0	4.0			51.6	9
B13	49	−4.0	16.0			46.8	14
B14	28	−25.0	625.0			30.2	20
B15	61	8.0	64.0			56.3	6
B16	42	−11.0	121.0			41.3	16
B17	52	−1.0	1.0			49.2	12
B18	54	1.0	1.0			50.8	10
B19	56	3.0	9.0			52.4	7
B20	79	26.0	676.0	160.0	12.6	70.6	1
合計	1060	0.0	3200.0			50.0	
個数	20		160.0				
最小値	28.0						
最大値	79.0						
中央値	54.0						
平均値	53.0		160.0		12.6		
最頻値	54.0						
分散	160.0						
標準偏差	12.6						
区間・階級	頻度						
9.9	0						
19.9	0						
29.9	1						
39.9	2						
49.9	4						
59.9	7						
69.9	4						
79.9	2						
89.9	0						
99.9	0						
100.9	0						
最小値	28						
第1四分位点	44.25						
中央値	54						
第3四分位点	61.75						
最大値	79						

ヒストグラム

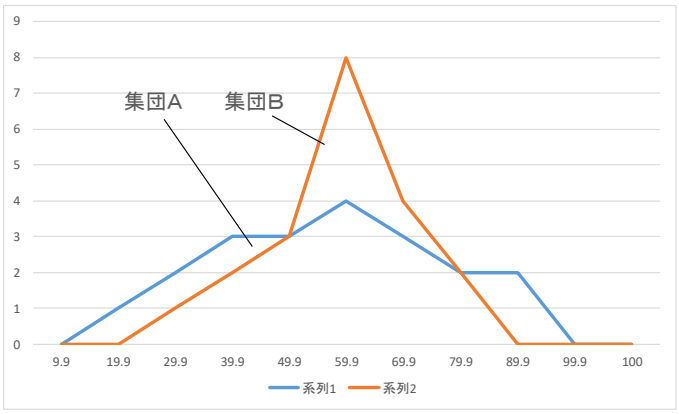
縦軸に度数、横軸に階級をとった統計グラフの一種で、データの分布状況を視覚的に認識するために主に統計学や数学、画像処理等で用いられる。

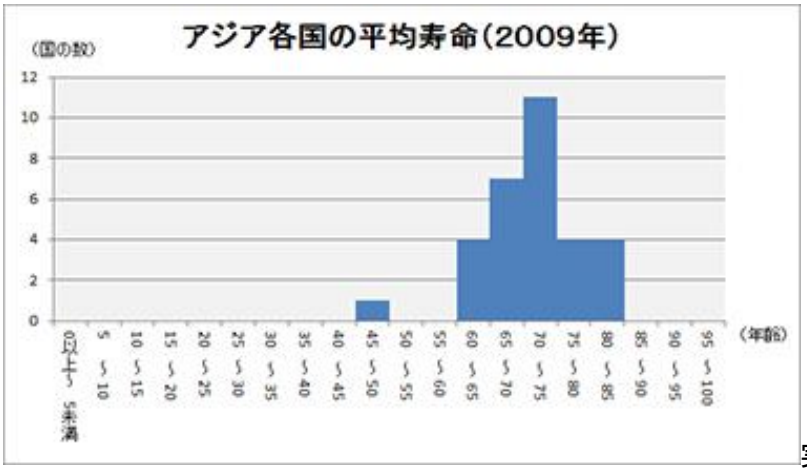


分散 379.2　標準偏差 19.5



分散 172.1　標準偏差 13.1

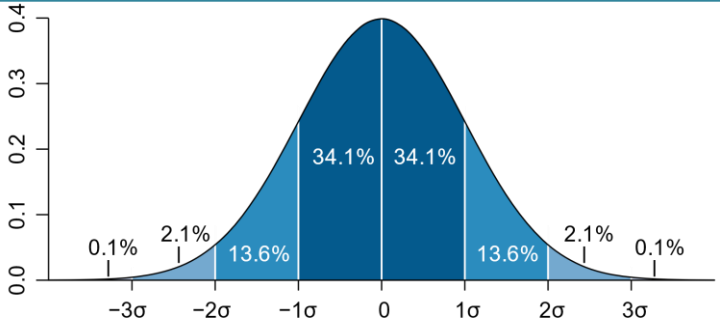




実例

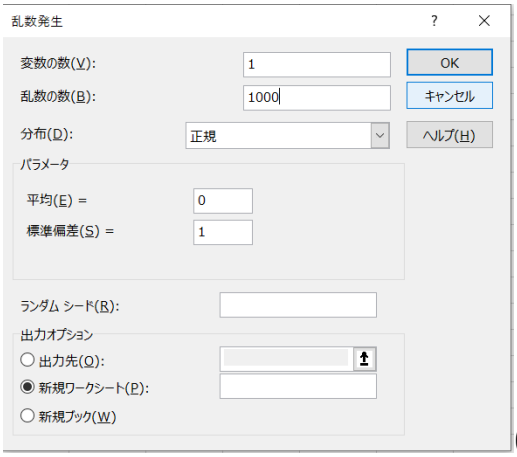
一般的に、テスト得点や身長の分布は、中央が一番高く両側に向かってだんだん低くなり左右対称の釣鐘型になる。この分布の型が「正規分布」であり、グラフは中央の一番高い位置が平均値。正規分布は、確率論や統計学で用いられる連続的な変数に関する確率分布の一つ。

(平均値±標準偏差×1)の範囲：全体の 68%の中にデータがある
(平均値±標準偏差×2)の範囲：全体の 95%の中にデータがある

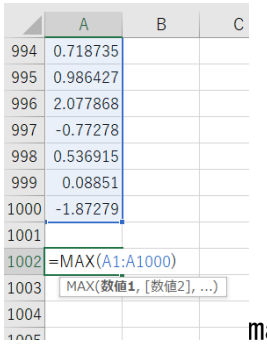


正規分布実験

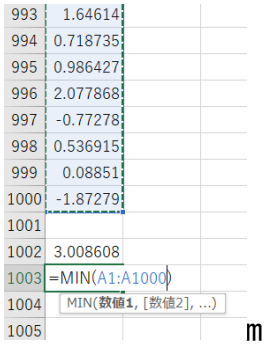
EXCEL を開き、新規ファイル➡データ➡データ分析➡分析ツール➡乱数発生➡OK



OK

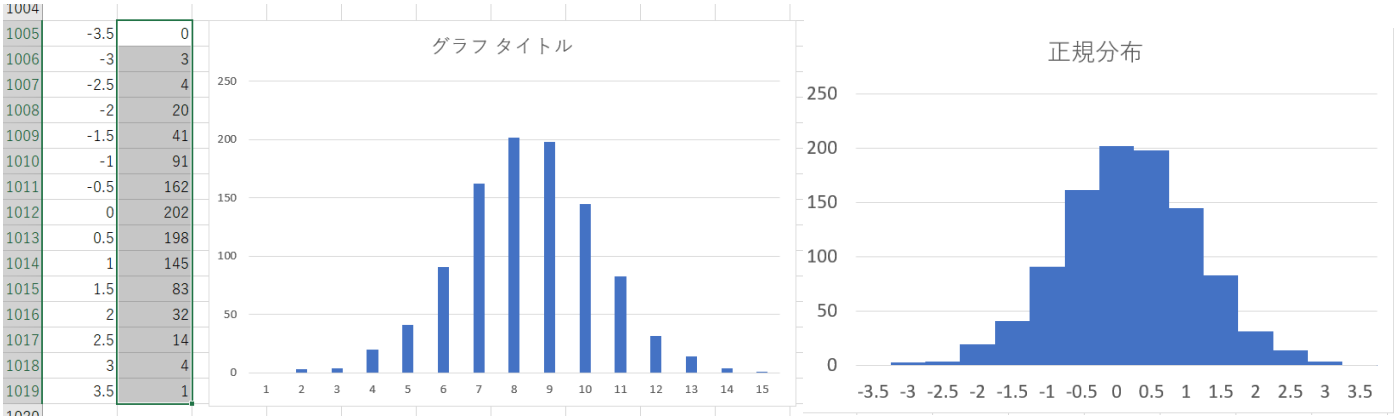


max



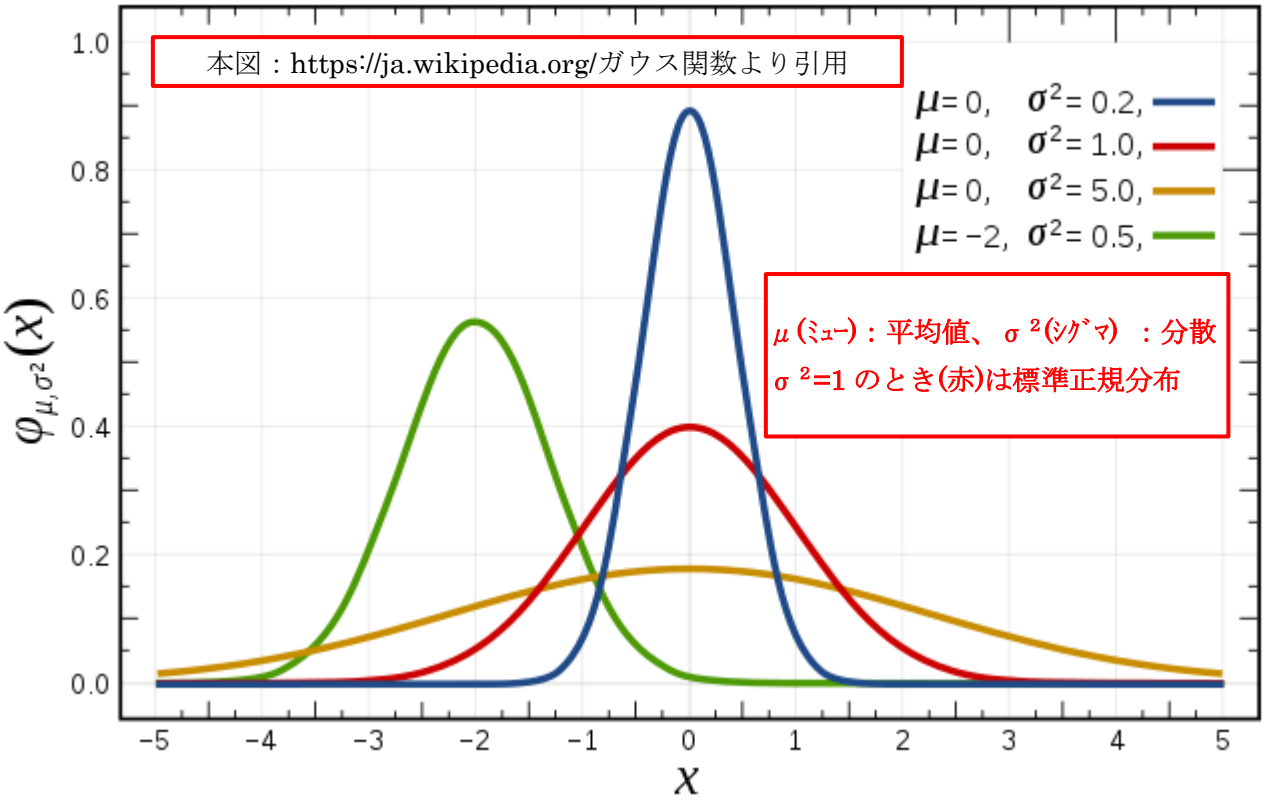
min

度数分布➡ヒストグラム



サイコロ 2 個を振ったときの出目の和➡確率変数 X に対する分布 P(X)

X	2	3	4	5	6	7	8	9	10	11	12
P(X)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36



以下の立式は数学にて

正規分布関数 $\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$

μ=0 ， 分散σ^2=1 の場合、標準正規分布 $\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$

偏差値 ＝｛（データ−平均値）／標準偏差｝×10+50　　正規分布が前提

偏差値 ＝「50±10」の間➡全体の 68.2%　偏差値 60➡集団の上位 15.8%

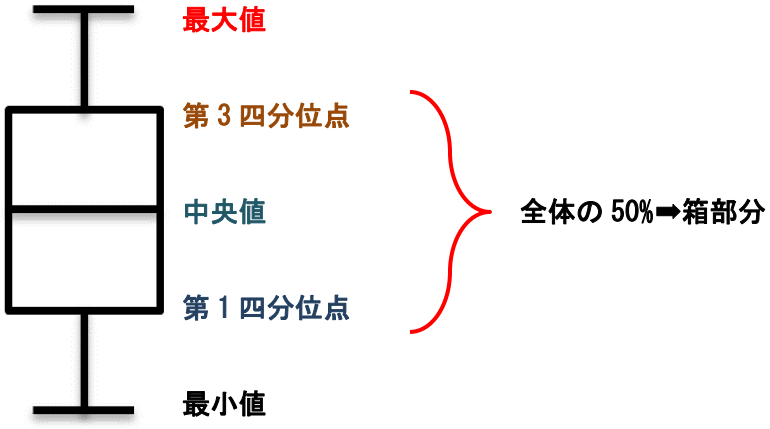
偏差値 ＝「50±20」の間➡全体の 95.6%　偏差値 70➡集団の上位 2.2%

偏差値 ＝「50±30」の間➡全体の 99.8%　偏差値 80➡集団の上位 0.1%

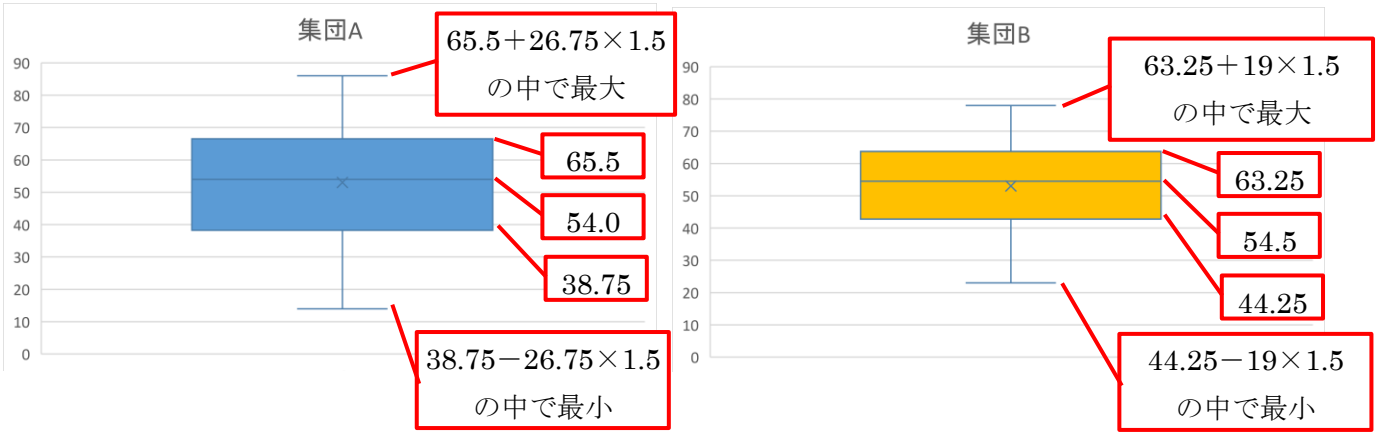
箱ひげ図

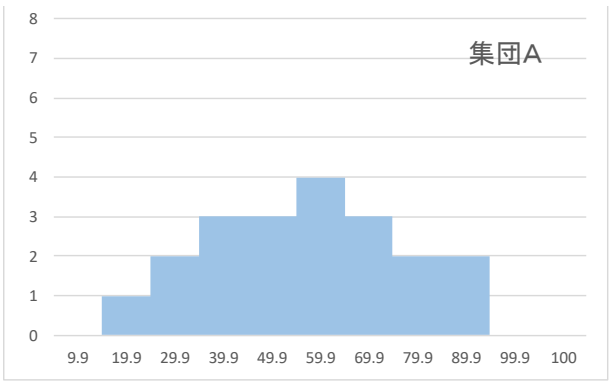
データのばらつきを視覚的に表現する統計図。主に多くの水準からなる分布を視覚的に要約し、比較可能

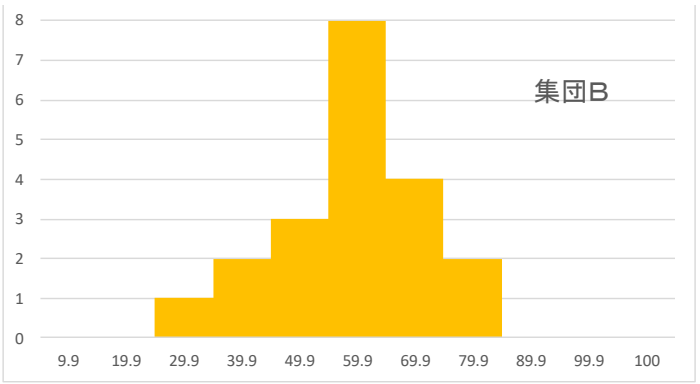
Q4/4 ： 最大値（maximum）
Q3/4 ： 第 3 四分位点（upper quartile）データの上位 25%
Q2/4 ： 中央値（第 2 四分位点、median）
Q1/4 ： 第 1 四分位点（lower quartile）データの下位 25%
Q0/4 ： 最小値（minimum）



最大値、第 3 四分位点、中央値（第 2 四分位点）、第 1 四分位点、最小値の統計量を表すグラフである。第 1 四分位点から第 3 四分位点までの高さに箱を描き、中央値で仕切りを描く。箱の上と下に「ひげ」を描く。ひげの長さは、箱の高さ（四分位範囲）の 1.5 倍以下の範囲にあるデータの中で、上端は最も大きいデータまで、下端は最も小さいデータまでとする。



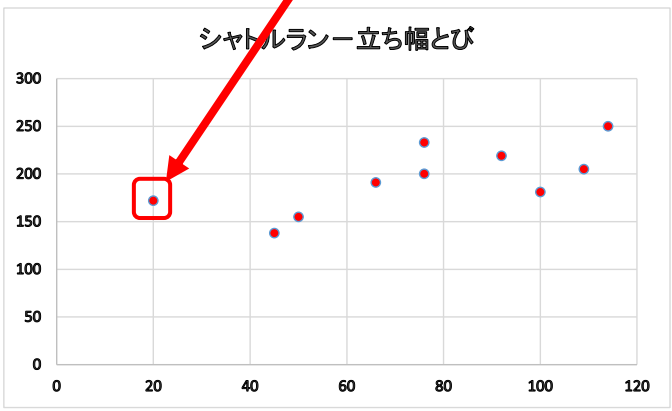


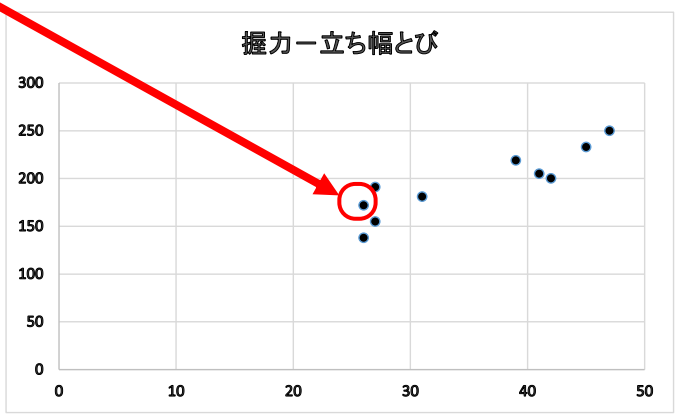


散布図　縦軸、横軸に 2 項目の量や大きさ等を対応させ、データを点でプロットした分布を示す

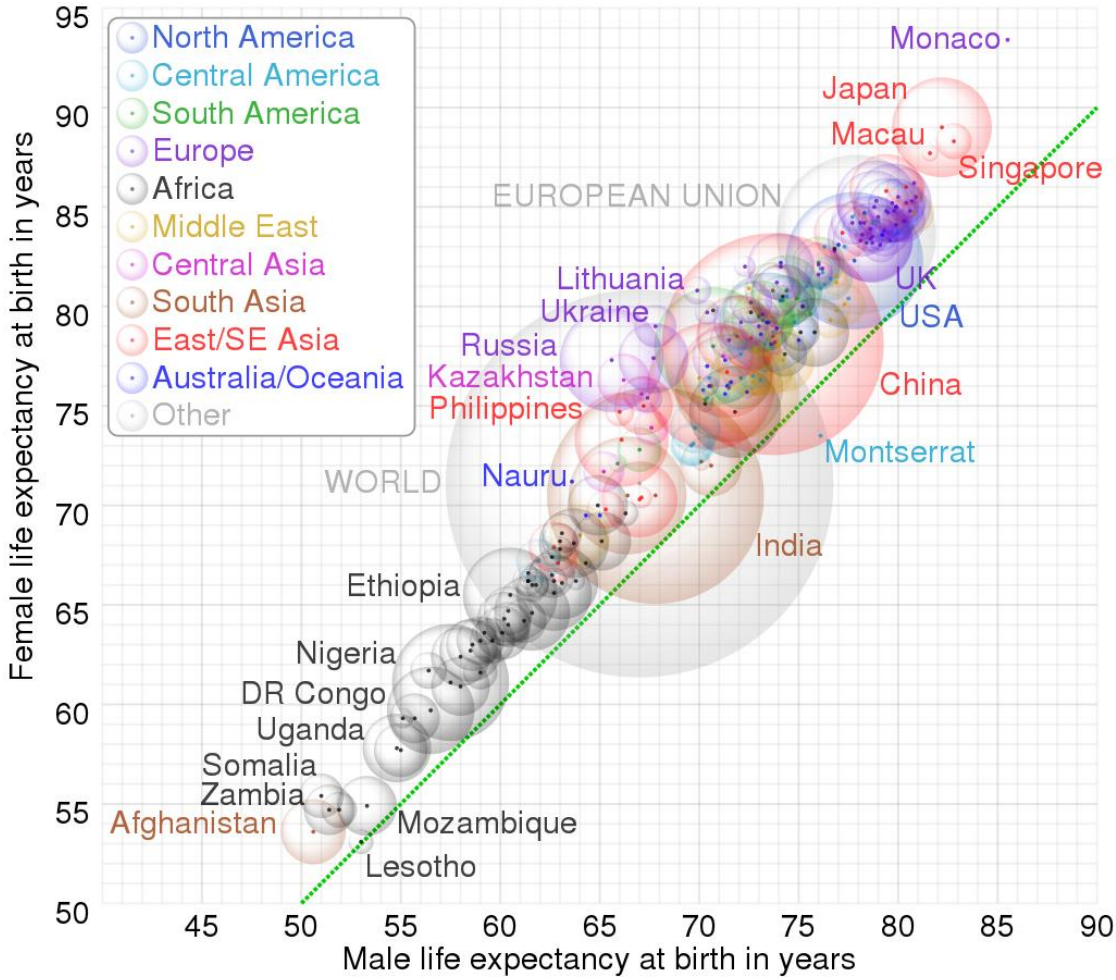
スポーツテスト・データの一部から

握力	20mシャトルラン	立ち幅とび
39	92	219
27	50	155
42	76	200
27	66	191
26	20	172
26	45	138
47	114	250
41	109	205
31	100	181
45	76	233





実例　各国の平均寿命の比較（出典：CIA ファクトブック、2018）



5-2　スポーツテスト結果分析

5-2-1　ＤＢ実験 4-6④により作成したテーブル SPT_T.xlsx を表計算ソフトウェアで開く

	A	B	C	D	E	F	G	H	I	J	K
1	No	grade	sex	name	club	20mShuttleRur	50mTime	BallThrow_dist	GripStrengt	LongJumpDi	SideJump
2	1	1	Male	t001	野球	92	7.8	21	39	219	53
3	2	1	Female	t002	その他	50	9.2	10	27	155	43
4	3	1	Male	t003	硬テニス	76	8.1	11	42	200	58
5	4	1	Female	t004	バスケ	66	8.7	13	27	191	52
6	5	1	Female	t005	なし	20	10.2	10	26	172	43
7	6	1	Female	t006	なし	45	9.8	8	26	138	46
8	7	1	Male	t007	水泳	114	7.6	27	47	250	64
9	8	1	Male	t008	軟テニス	109	7.8	30	41	205	63
10	9	1	Female	t009	陸上-走	100	8.1	16	31	181	53
11	10	1	Male	t010	陸上-走	76	7.1	29	45	233	57
12	11	1	Female	t011	なし	45	8.9	7	23	170	50
13	12	1	Male	t012	陸上-走	114	6.6	27	47	255	65
14	13	1	Female	t013	バスケ	54	9.3	12	25	140	48
15	14	1	Male	t014	サッカー	110	7.2	20	32	228	61
16	15	1	Female	t015	なし	35	9.7	10	25	175	48
17	16	1	Female	t016	その他	72	8.5	19	29	190	60
18	17	1	Female	t017	その他	34	9.6	9	25	156	44
19	18	1	Female	t018	硬テニス	78	8.7	13	24	170	52
20	19	1	Male	t019	硬テニス	90	7.3	20	36	223	56
21	20	1	Female	t020	バレーボール	90	8.8	17	40	192	60
22	21	1	Female	t021	なし	48	9.6	8	25	159	45
					↕802 人						
797	796	3	Male	t796	軟テニス	113	7.4	26	49	222	63
798	797	3	Male	t797	山岳	72	7.8	20	42	210	39
799	798	3	Male	t798	水泳	125	7.8	18	50	228	62
800	799	3	Female	t799	なし	51	9.3	9	28	160	48
801	800	3	Female	t800	陸上-投	40	8.1	26	34	210	57
802	801	3	Male	t801	野球	125	7.1	22	55	240	63
803	802	3	Female	t802	バドミントン	79	9.2	13	25	157	56
804											

5-2-2　各種目の代表値を所定のセルに求めよ。

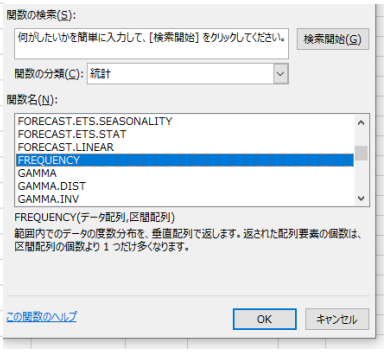
		20mSRun	50m_Time	Ball throw	Grip strength	LongJump	Side jump
810							
811	合計	=SUM(F2:F803)	=SUM(G2:G803)	=SUM(H2:H803)	=SUM(I2:I803)	=SUM(J2:J803)	=SUM(K2:K803)
812	個数	=COUNT(F2:F803)	=COUNT(G2:G803)	=COUNT(H2:H803)	=COUNT(I2:I803)	=COUNT(J2:J803)	=COUNT(K2:K803)
813	最小値	=MIN(F2:F803)	=MIN(G2:G803)	=MIN(H2:H803)	=MIN(I2:I803)	=MIN(J2:J803)	=MIN(K2:K803)
814	最大値	=MAX(F2:F803)	=MAX(G2:G803)	=MAX(H2:H803)	=MAX(I2:I803)	=MAX(J2:J803)	=MAX(K2:K803)
815	中央値	=MEDIAN((F2:F803))	=MEDIAN((G2:G803))	=MEDIAN((H2:H803))	=MEDIAN((I2:I803))	=MEDIAN((J2:J803))	=MEDIAN((K2:K803))
816	平均値	=AVERAGE(F2:F803)	=AVERAGE(G2:G803)	=AVERAGE(H2:H803)	=AVERAGE(I2:I803)	=AVERAGE(J2:J803)	=AVERAGE(K2:K803)
817	最頻値	=MODE.SNGL((F2:F803))	=MODE.SNGL((G2:G803))	=MODE.SNGL((H2:H803))	=MODE.SNGL((I2:I803))	=MODE.SNGL((J2:J803))	=MODE.SNGL((K2:K803))
818	分散	=VAR.P((F2:F803))	=VAR.P((G2:G803))	=VAR.P((H2:H803))	=VAR.P((I2:I803))	=VAR.P((J2:J803))	=VAR.P((K2:K803))
819	標準偏差	=STDEV.P((F2:F803))	=STDEV.P((G2:G803))	=STDEV.P((H2:H803))	=STDEV.P((I2:I803))	=STDEV.P((J2:J803))	=STDEV.P((K2:K803))

	20mSRun	50m_Time	Ball throw	Grip strengt	LongJump	Side jump
合計	56760	6558.88	13632	26942	152566	41894
個数	773	778	797	802	789	786
最小値	1	6.4	5	17	70	23
最大値	127	11.8	39	61	334	69
中央値	71	9	16	32	192	53
平均値	73	8	17	34	193	53
最頻値	125	9	11	27	180	53
分散	977	1	50	88	1200	53
標準偏差	31	1	7	9	35	7

頻度の算出

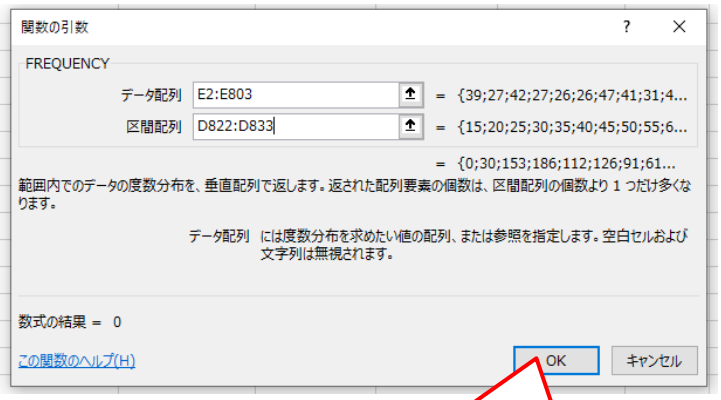
①　値を返すセルを指定　②数式➡関数の挿入➡統計➡frequency

区間	頻度
15	
20	
25	
30	
35	
40	
45	
50	
55	
60	
65	
70	



③　右下の「作法」で実行

区間	頻度
15	D833)
20	186
25	112
30	126
35	91
40	61
45	33
50	9
55	1
60	0
65	0
70	0

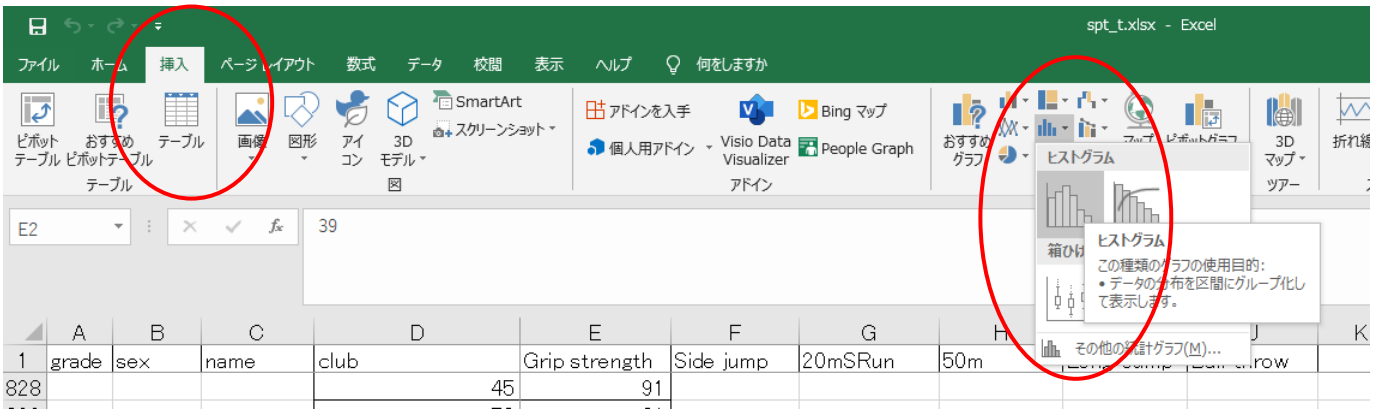


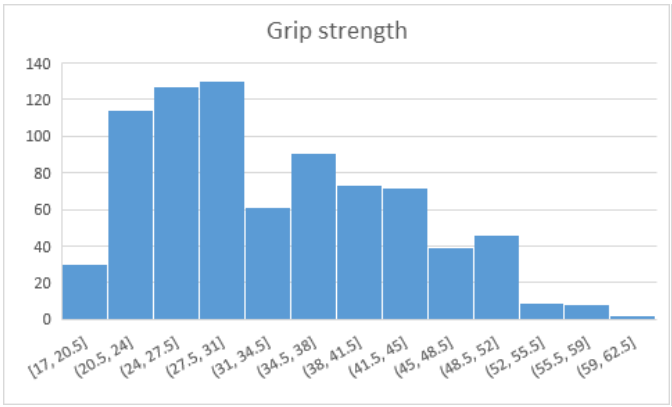
Shift、Ctrl を押しながら

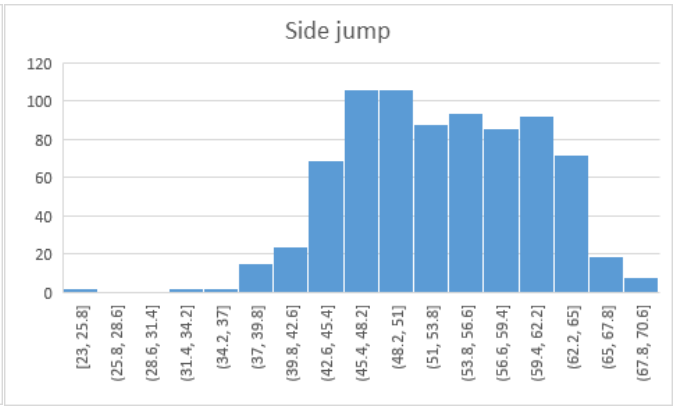
解答シート　sports test_answer　を開くパスワードは授業にて

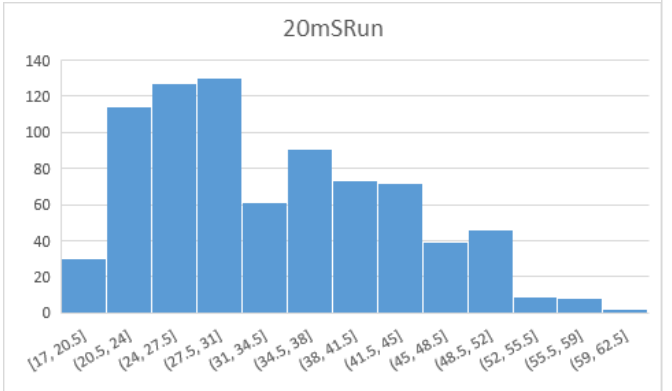
④　種目を選択し、ヒストグラムを描く。　➡「棒グラフ」とは異なる

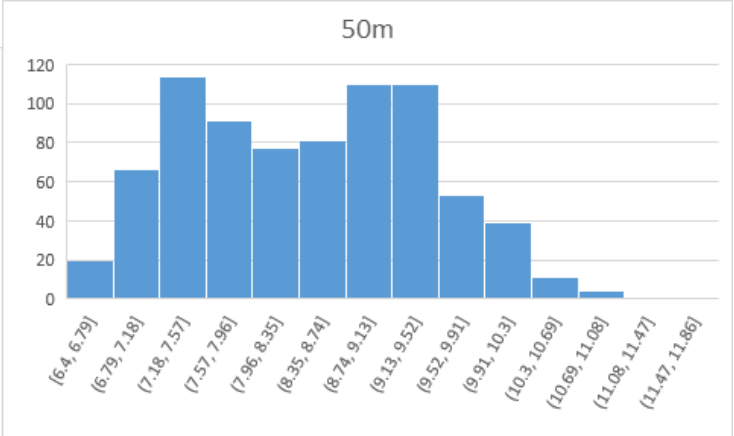
E2:E803 を範囲指定➡挿入➡ヒストグラム

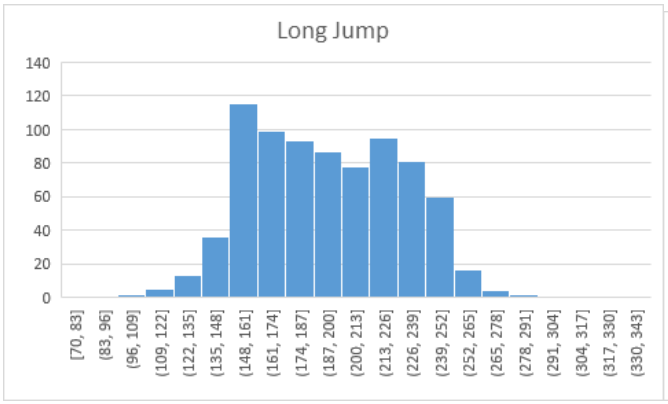


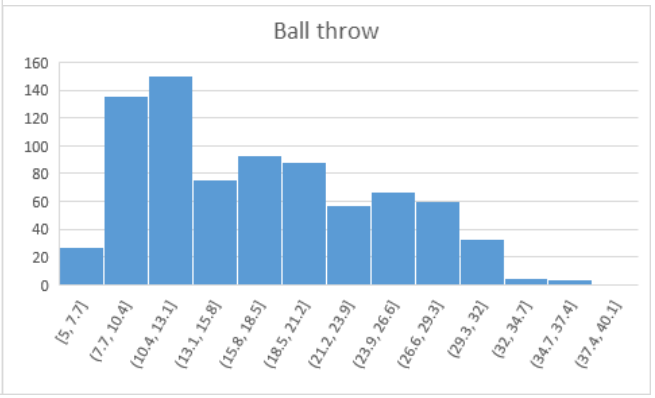






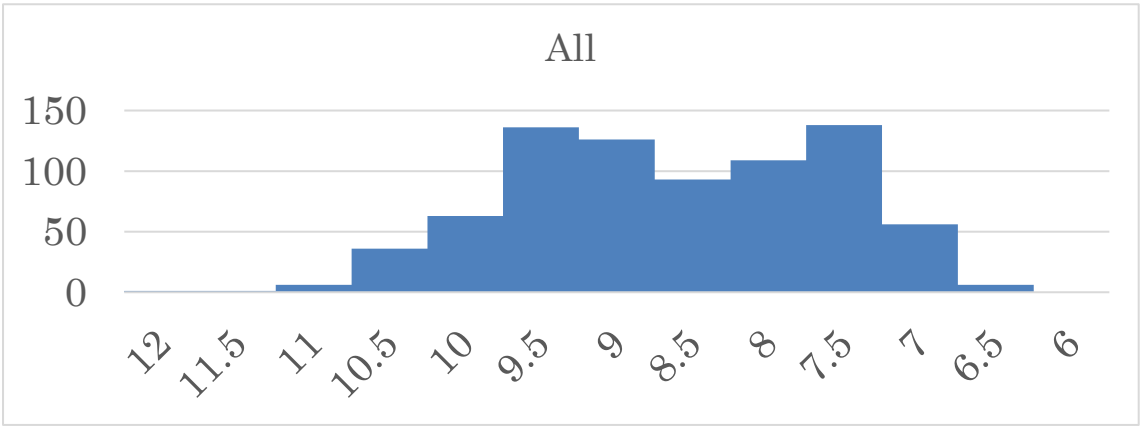






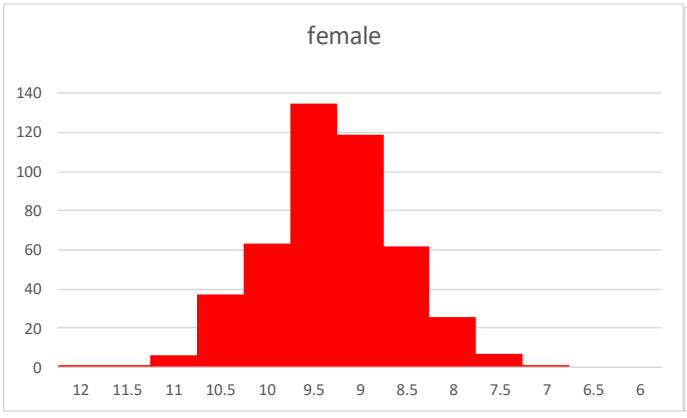
考察：6 種目ともきれいな正規分布とは言えない。理由を推測し、データを基に実証しよう。

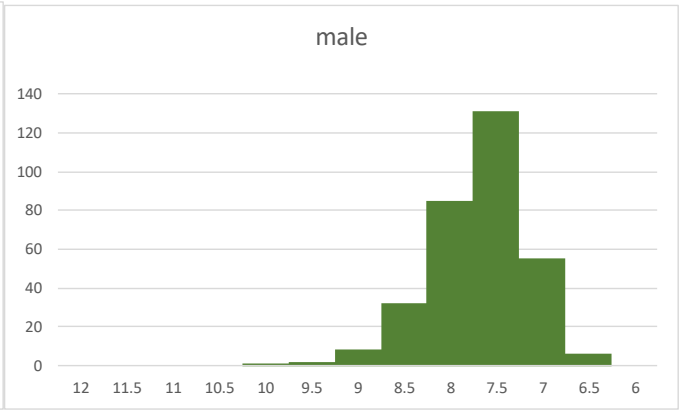
例　50m 走 Time のヒストグラム



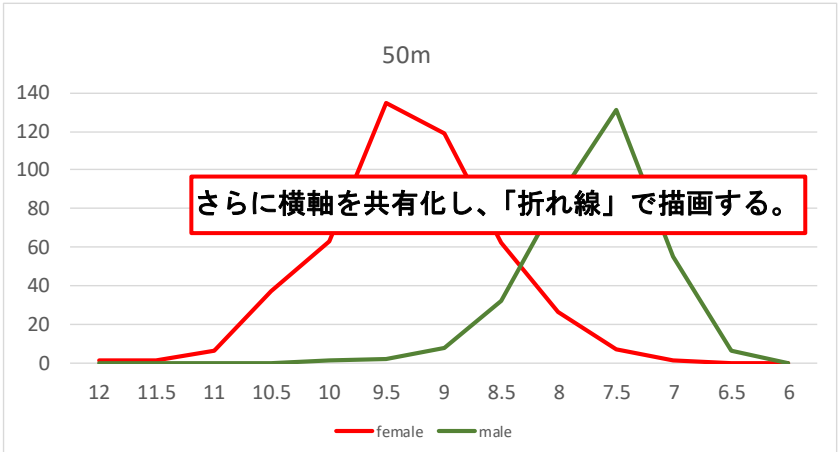
実証　例としてツインピークが顕著な 50m 走で「性別(sex)」に起因するものと仮定する。

表計算ソフトウェアで male, female でフィルタを実施し、個々に縦棒グラフからヒストグラムを描画する。



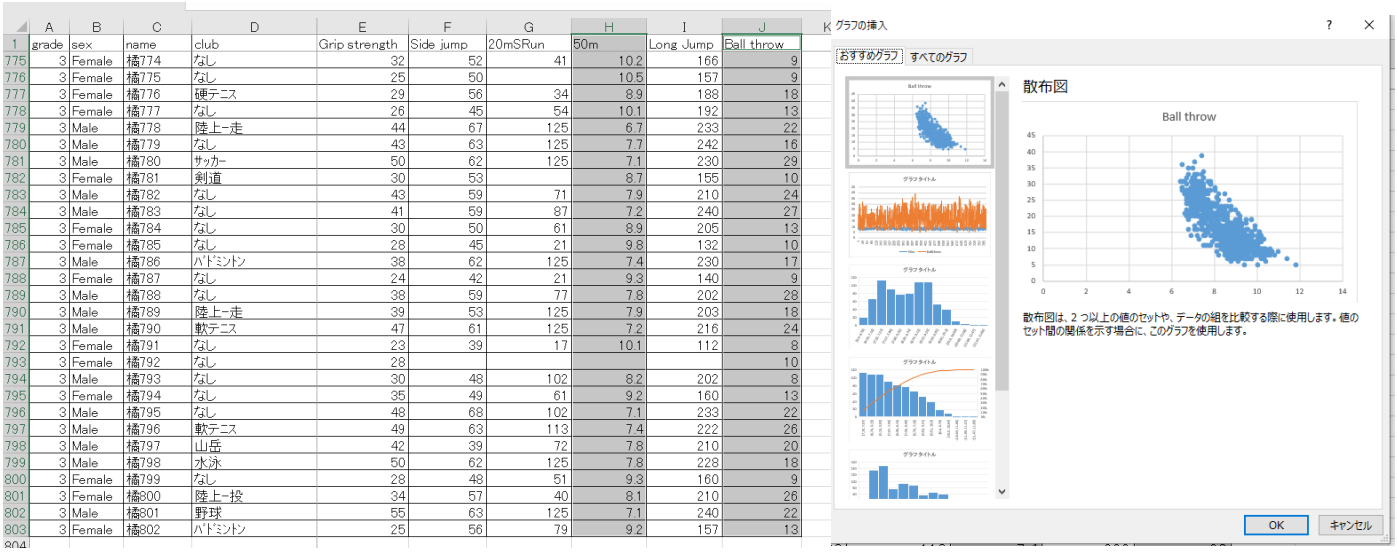


time	female	male
12	1	0
11.5	1	0
11	6	0
10.5	37	0
10	63	1
9.5	135	2
9	119	8
8.5	62	32
8	26	85
7.5	7	131
7	1	55
6.5	0	6
6	0	0

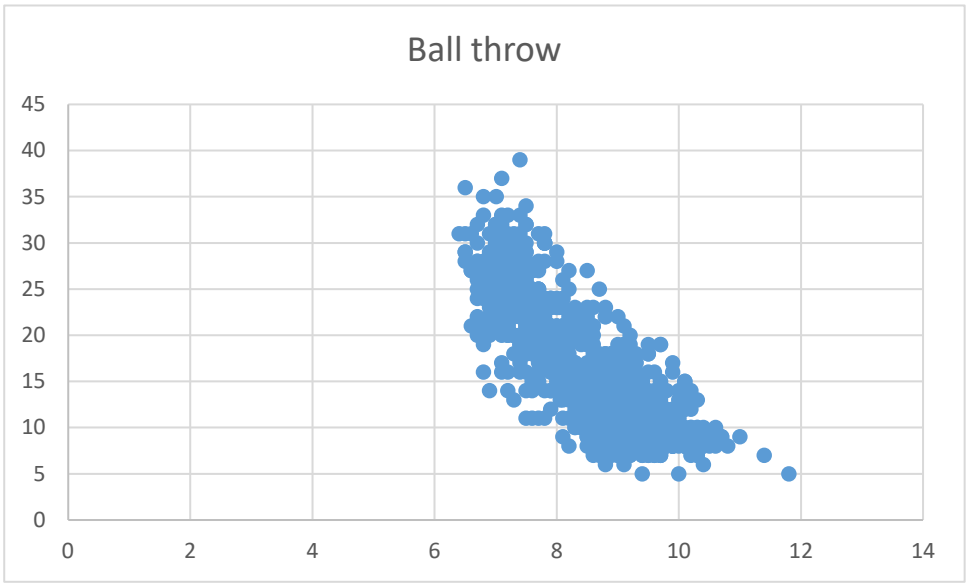


性別で分けた分析では、男女ともほぼ正規分布しており、これらが重なったヒストグラムが二峰性になった。

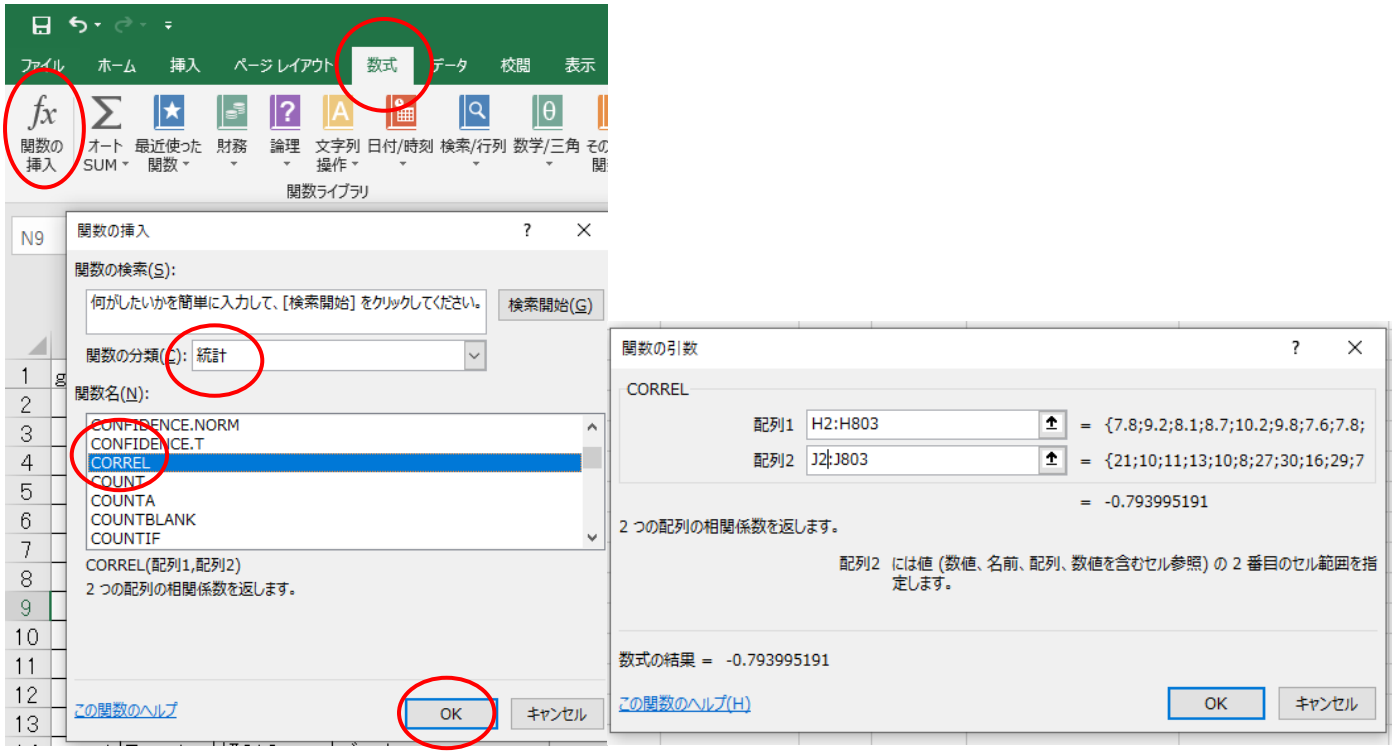
⑤　関連性が強い（弱い）と考えられる 2 項目を予測し、分布図を描く。2 列を選択➡挿入➡グラフ➡散布図



	A	B	C	D	E	F	G	H	I	J
1	grade	sex	name	club	Grip strength	Side jump	20mSRun	50m	Long Jump	Ball throw
775	3	Female	橘774	なし	32	52	41	10.2	166	9
776	3	Female	橘775	なし	25	50		10.5	157	9
777	3	Female	橘776	硬テニス	29	56	34	8.9	188	18
778	3	Female	橘777	なし	26	45	54	10.1	192	13
779	3	Male	橘778	陸上-走	44	67	125	6.7	233	22
780	3	Male	橘779	なし	43	63	125	7.7	242	16
781	3	Male	橘780	サッカー	50	62	125	7.1	230	29
782	3	Female	橘781	剣道	30	53		8.7	155	10
783	3	Male	橘782	なし	43	59	71	7.9	210	24
784	3	Male	橘783	なし	41	59	87	7.2	240	27
785	3	Female	橘784	なし	30	50	61	8.9	205	13
786	3	Female	橘785	なし	28	45	21	9.8	132	10
787	3	Male	橘786	ﾊﾞﾄﾞﾐﾝﾄﾝ	38	62	125	7.4	230	17
788	3	Female	橘787	なし	24	42	21	9.3	140	9
789	3	Male	橘788	なし	38	59	77	7.8	202	28
790	3	Male	橘789	陸上-走	39	53	125	7.9	203	18
791	3	Male	橘790	軟テニス	47	61	125	7.2	216	24
792	3	Female	橘791	なし	23	39	17	10.1	112	8
793	3	Female	橘792	なし	28					10
794	3	Male	橘793	なし	30	48	102	8.2	202	8
795	3	Female	橘794	なし	35	49	61	9.2	160	13
796	3	Male	橘795	なし	48	68	102	7.1	233	22
797	3	Male	橘796	軟テニス	49	63	113	7.4	222	26
798	3	Male	橘797	山岳	42	39	72	7.8	210	20
799	3	Male	橘798	水泳	50	62	125	7.8	228	18
800	3	Female	橘799	なし	28	48	51	9.3	160	9
801	3	Female	橘800	陸上-投	34	57	40	8.1	210	26
802	3	Male	橘801	野球	55	63	125	7.1	240	22
803	3	Female	橘802	ﾊﾞﾄﾞﾐﾝﾄﾝ	25	56	79	9.2	157	13



⑥　相関係数を関数で求める　　解を返すセルを指定➡数式➡関数挿入➡統計➡CORREL➡OK➡2 列選択



相関係数	20mShuttleRun	50mTime	BallThrow_distance	GripStrength	LongJumpDistance	SideJump
20mShuttleRun	–	=CORREL($F$2:$F$803,G$2:G$803)	=CORREL($F$2:$F$803,H$2:H$803)	=CORREL($F$2:$F$803,I$2:I$803)	=CORREL($F$2:$F$803,J$2:J$803)	=CORREL($F$2:$F$803,K$2:K$803)
50mTime	–	–	=CORREL($G$2:$G$803,H$2:H$803)	=CORREL($G$2:$G$803,I$2:I$803)	=CORREL($G$2:$G$803,J$2:J$803)	=CORREL($G$2:$G$803,K$2:K$803)
BallThrow_distance	–	–	–	=CORREL($H$2:$H$803,I$2:I$803)	=CORREL($H$2:$H$803,J$2:J$803)	=CORREL($H$2:$H$803,K$2:K$803)
GripStrength	–	–	–	–	=CORREL($I$2:$I$803,J$2:J$803)	=CORREL($I$2:$I$803,K$2:K$803)
LongJumpDistance	–	–	–	–	–	=CORREL(J2:J803,K2:K803)
SideJump	–	–	–	–	–	–

相関係数	20mShuttleRu	50mTime	BallThrow_dis	GripStrengt	LongJumpD	SideJump
20mShuttleRun	—	−0.79	0.70	0.66	0.75	0.76
50mTime	—	—	−0.79	−0.79	−0.87	−0.76
BallThrow_distance	—	—	—	0.79	0.79	0.72
GripStrength	—	—	—	—	0.76	0.67
LongJumpDistance	—	—	—	—	—	0.74
SideJump	—	—	—	—	—	—

相関係数 r の値	相関
−1≤r≤−0.7	強い負の相関
−0.7≤r≤−0.4	負の相関
−0.4≤r≤−0.2	弱い負の相関
−0.2≤r≤0.2	ほとんど相関がない
0.2≤r≤0.4	弱い正の相関
0.4≤r≤0.7	正の相関
0.7≤r≤1	強い正の相関

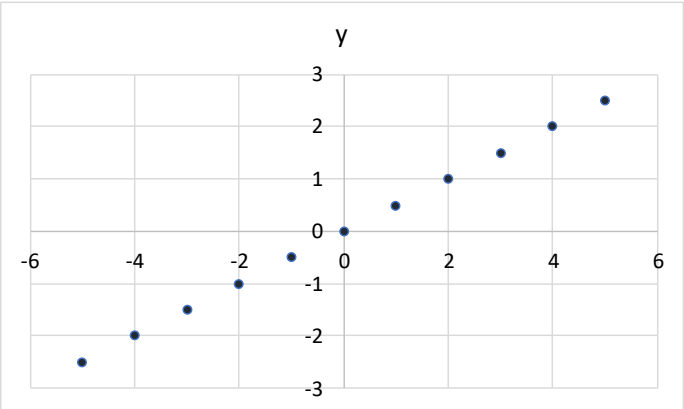
➡小実験「正比例」

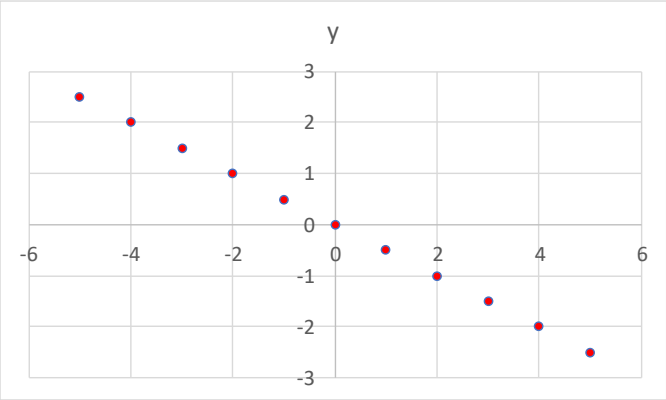
x	y
-5	-2.5
-4	-2
-3	-1.5
-2	-1
-1	-0.5
0	0
1	0.5
2	1
3	1.5
4	2
5	2.5

相関係数 1

x	y
-5	2.5
-4	2
-3	1.5
-2	1
-1	0.5
0	0
1	-0.5
2	-1
3	-1.5
4	-2
5	-2.5

相関係数 -1





注意点：相関と因果は似て非なる概念である。相関は「木の高さと幹の太さ」のように関連性はあっても順序を問わないケースである。「両掌を擦ると熱が出る」は因果であるとともに、強い関連性（相関）もある。因果関係とは、ある出来事（擦る）が別の出来事（まさつ熱）を直接的に引き起こす原因と結果の関係。逆に手のひらを温めても擦る運動にはならない。相関は因果の単なる必要条件の 1 つ従って本件において 50m 走タイムとボール投げ距離の間に強い関連性（相関）は認められるが、50m 走が速く走れることがボールを遠くに投げる要因とは言えない。

虚偽の原因：A の発生は B と相関している。したがって、A が B の原因である。

様々なケース

・**理想気体の状態方程式 PV=nRT は圧力と温度の関係**を示し、両者に相関がある。➡質量が変わらない場合圧力を高くすると温度が上がり、温度を高くすると圧力が上がる比例関係にある。

・**アイスクリームの売り上げが伸びると、水難事故も確実に増える。よって、アイスクリームが水死の原因**だ。➡夏の暑さが両方の事象の共通する原因

・**1950 年代以降、大気の CO_2 レベルと犯罪レベルは同時に増大してきた。よって、CO_2 増加が犯罪増加の原因だ。**➡

・**明かりをつけたまま眠る若者は、その後近視になる可能性が高い。**

➡乳児を明かりをつけたまま寝かせることと近視に関係があるという結果は得られなかった。一方、親が近視の子供は近視になる確率が高いという結果が得られ、近視の親が子供を明かりをつけた寝室で寝かせることが多いという傾向があった。交絡変数は、両親の近視と考えられる

⑦　各種目の箱ひげ図を描画

⑦-1 先頭行を指定

	A	B	C	D	E	F	G	H	I	J
1	grade	sex	name	club	Grip strength	Side jump	20mSRun	50m	Long Jump	Ball throw
2	1	Male	橘001	野球	39	53	92	7.8	219	21
3	1	Female	橘002	その他	27	43	50	9.2	155	10
4	1	Male	橘003	硬テニス	42	58	76	8.1	200	11

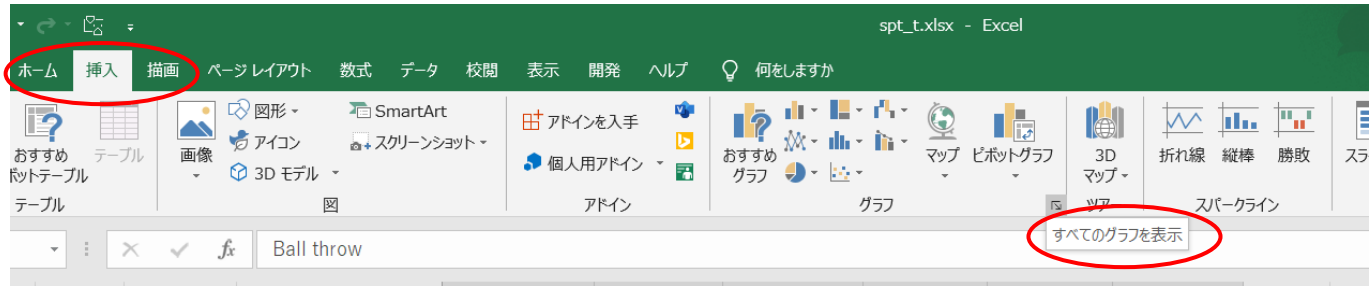
⑦-2 Shift キーを押しながら最終行を指定

	A	B	C	D	E	F	G	H	I	J
1	grade	sex	name	club	Grip strength	Side jump	20mSRun	50m	Long Jump	Ball throw
799	3	Male	橘798	水泳	50	62	125	7.8	228	18
800	3	Female	橘799	なし	28	48	51	9.3	160	9
801	3	Female	橘800	陸上-投	34	57	40	8.1	210	26
802	3	Male	橘801	野球	55	63	125	7.1	240	22
803	3	Female	橘802	ﾊﾞﾄﾞﾐﾝﾄﾝ	25	56	79	9.2	157	13
804										

⑦-3 Ctrl キーを押しながら 6 種目をフィールドごとに指定

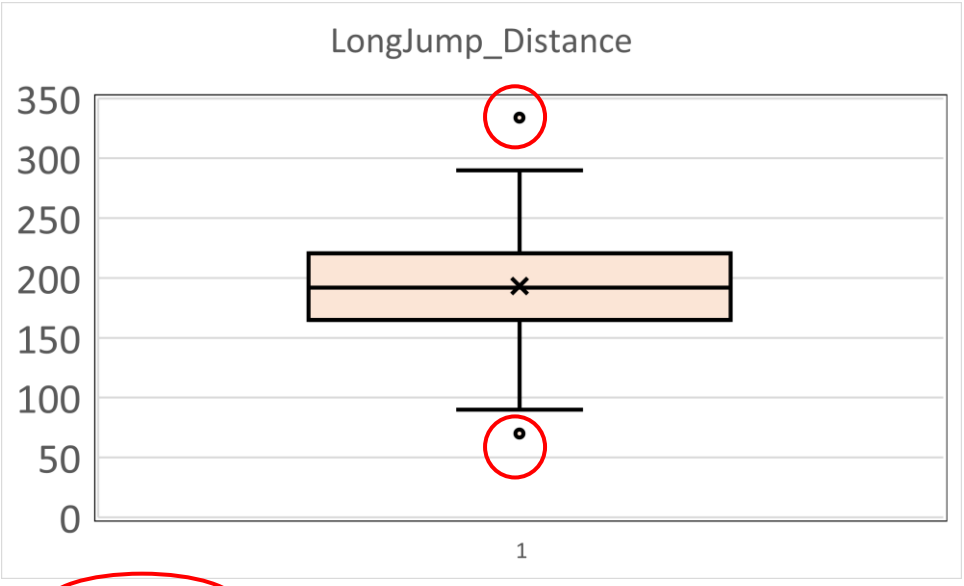
	A	B	C	D	E	F	G	H	I	J
1	grade	sex	name	club	Grip strength	Side jump	20mSRun	50m	Long Jump	Ball throw
785	3	Female	橘784	なし	30	50	61	8.9	205	13
786	3	Female	橘785	なし	28	45	21	9.8	132	10
787	3	Male	橘786	ﾊﾞﾄﾞﾐﾝﾄﾝ	38	62	125	7.4	230	17
788	3	Female	橘787	なし	24	42	21	9.3	140	9
789	3	Male	橘788	なし	38	59	77	7.8	202	28
790	3	Male	橘789	陸上-走	39	53	125	7.9	203	18
791	3	Male	橘790	軟テニス	47	61	125	7.2	216	24
792	3	Female	橘791	なし	23	39	17	10.1	112	8
793	3	Female	橘792	なし	28					10
794	3	Male	橘793	なし	30	48	102	8.2	202	8
795	3	Female	橘794	なし	35	49	61	9.2	160	13
796	3	Male	橘795	なし	48	68	102	7.1	233	22
797	3	Male	橘796	軟テニス	49	63	113	7.4	222	26
798	3	Male	橘797	山岳	42	39	72	7.8	210	20
799	3	Male	橘798	水泳	50	62	125	7.8	228	18
800	3	Female	橘799	なし	28	48	51	9.3	160	9
801	3	Female	橘800	陸上-投	34	57	40	8.1	210	26
802	3	Male	橘801	野球	55	63	125	7.1	240	22
803	3	Female	橘802	ﾊﾞﾄﾞﾐﾝﾄﾝ	25	56	79	9.2	157	13
804										

⑦-4 メニューバー　挿入➡グラフ➡すべてのグラフ➡箱ひげ図を選択



以下、描画については授業にて

スポーツテスト種目別測定結果 802 人分の分布



に注目：外れ値➡データ分布において、他の観測値から大きく外れた値のこと。
測定ミスや異常を伴う観測など、様々な原因
ひげの下端より小さい値、ひげの上端より大きい値

コラム　仮説検定

統計的仮説検定：確率をもとに結論を導く方法。

①仮説を立てる　②結果を確率的に検証➡**背理法**により結論。

背理法：仮説を設定し、仮説が正しいとした条件で矛盾が生じた場合に仮説が間違っていると判断する方法

例：振った時、①　50%の確率で表が出る正常コイン 2 枚、②　10%の確率で表が出るイカサマコイン 2 枚が存在

➡2 枚セットを 2 回振った時に「裏裏」「裏裏」の結果であった➡正常コイン、イカサマコイン

1. 仮説：正常コインとする

2. 検定：「裏裏」「裏裏」の確率

正常コインの場合 $0.5^2 \times 0.5^2 = 0.0625 = 6.25\%$

3. 判断：　6.25%…very rare➡イカサマコイン　　または very rare だが起こりえる➡正常コイン

4. 結論：設定基準に基づく

(1) 設定基準：10%

➡事象の評価は 10%未満の確立なら仮説が誤り、10%以上の確立なら許容範囲で仮説が誤りといえない」と判断

本事例では、仮説を正常コインとしたが、6.25%は判断基準 10%未満➡イカサマコイン（絶対ではない）

注意点：導かれた結論は「絶対」と考えることはできない「正常コインでも、6.25%起こりえる事象

(2) 設定基準：5%➡仮説：正常コインは誤りではない