ゼロからはじめる統計学

第4回

「統計3級範囲」



確率

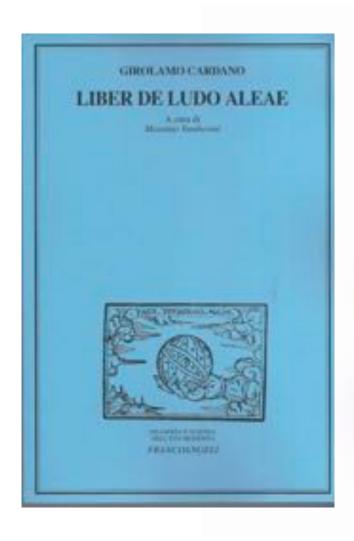
和から株式会社

確率とは?



ジロラモ・カルダノ 1501~1576

著書『さいころあそびについて』



ギャンブラーにとっては、 全くギャンブルをしないことが 最大の利益となる。

和から株式会社

科学革命と確率



ブレーズ・パスカル 1623~1662

アカデミー・メルセンヌ



マラン・メルセンヌ



ルネ・デカルト



ピエール・ド・フェルマー



ブレーズ・パスカル (当時10代)

問題:先に10回勝つのはどちらか?





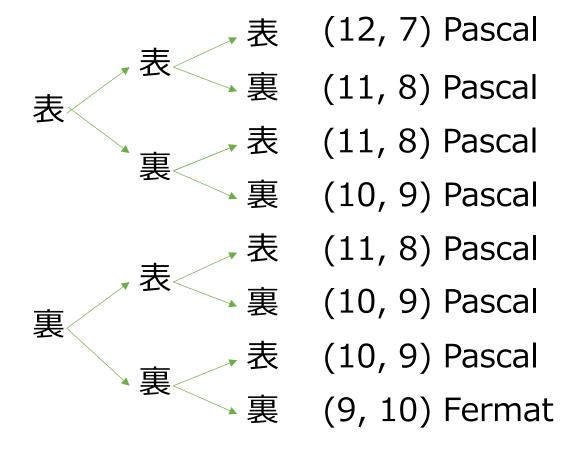


Pascal	Fermat
9	7

中断した時、賞金の配分はどうすれば良いか?

問題:先に10回勝つのはどちらか?

全ての可能性を調べる



問題:先に10回勝つのはどちらか?

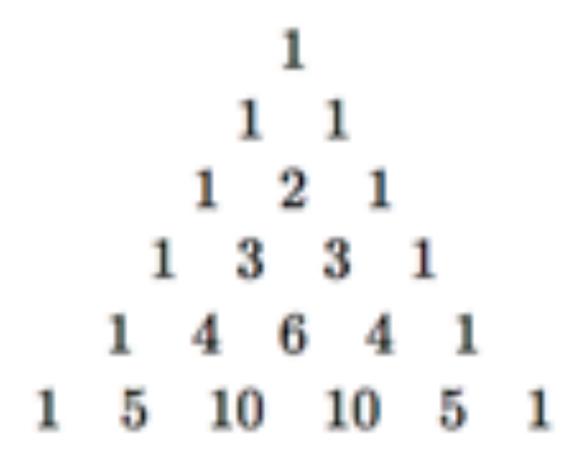


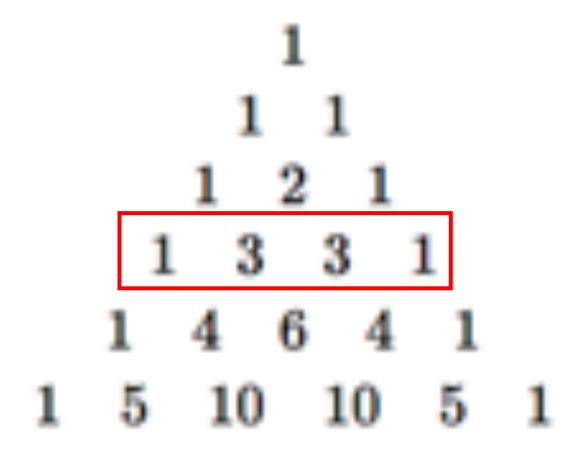
Pascal	Fermat
9	7

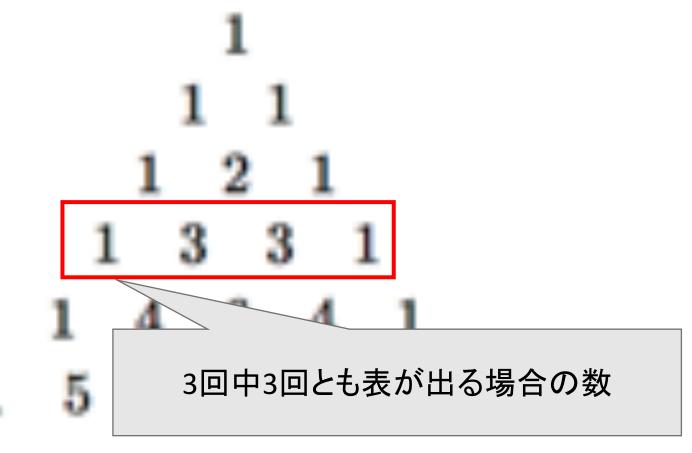


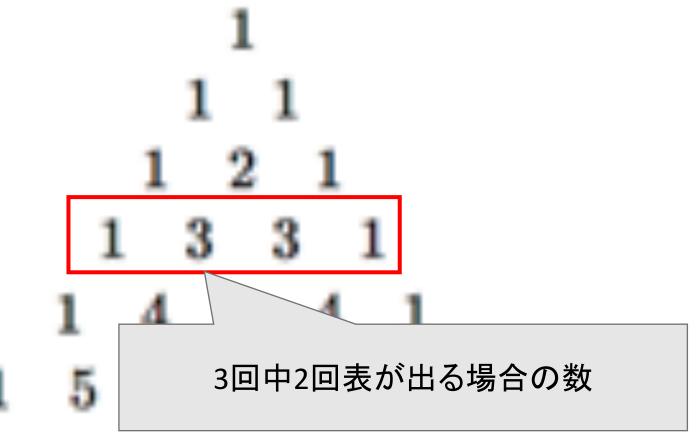
勝率

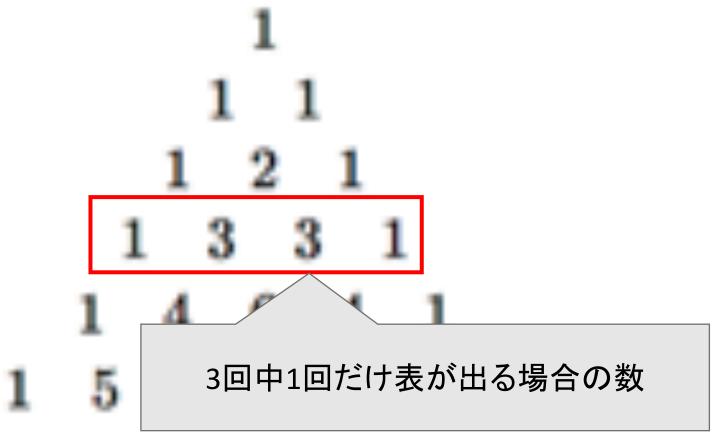
Pascal	Fermat
7/8	1/8
(87.5%)	(12.5%)

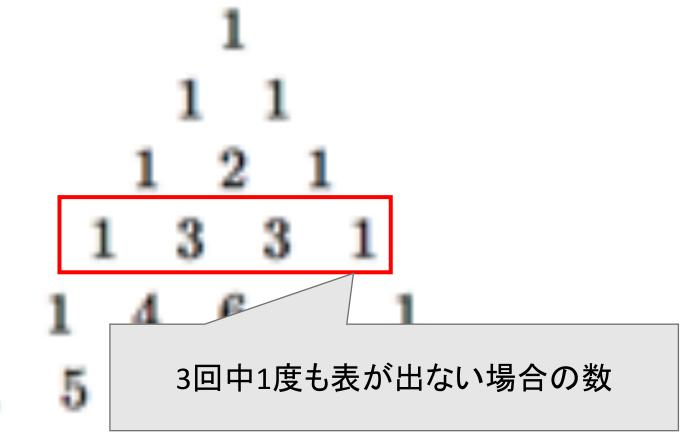












確率

起こりうる結果がn通りあり、そのうちのk通りある出来事Aが起こる確率P(A)は:

$$P(A) = \frac{k}{n}$$

例:宝くじが当たる確率は?

ドリームジャンボ宝くじ 1億3000万枚 一等13本

13/1.3億 = 0.0000001 (飛行機が落ちる確率と同程度)





条件付き確率

ある家庭には双子の子供がいる。 その家を訪ねたところ、玄関に女児用のサンダルを 一足だけ確認できた。 この時、双子が二人とも女の子である確率はいくらか?



条件付き確率



(男、女)

(女、男)

(女、女)



 $\frac{1}{3}$

事例:コリンズ裁判

窃盗事件の目撃証言

部分的に黄色い自動車 … 0.1 口ひげをつけている … 0.25 あごひげをつけた黒人 … 0.1 ポニーテールの女性 … 0.1 ブロンドの女性 … 0.3 車の中に異人種のカップル … 0.001



これらの条件が全て揃う条件付き確率は… 0.1 × 0.25 × 0.1 × 0.1 × 0.3 × 0.001 = 約10億分の一!

犯人に間違いない!!

事例:コリンズ裁判

窃盗事件の目撃証言

部分的に黄色い自動車 … 0.1 口ひげをつけている … 0.25 あごひげをつけた黒人 … 0.1 ポニーテールの女性 … 0.1 ブロンドの女性 … 0.3 車の中に異人種のカップル … 0.001

口ひげがあればあごひげが ある可能性も高い

→独立ではない

実際にははるかに小さい確率…

演習問題:条件付き確率

高校生A君はT大学への進学を希望しています。 前期試験の合格率は60%です。 前期試験に不合格だった場合後期試験を受けられますが、 後期試験は定員が少ないため合格率は10%になります。 A君がT大学に合格する確率はいくらでしょうか?

演習問題:条件付き確率

前期試験に合格する場合

0.6 (60%)

前期試験に不合格となり、後期試験に合格する場合

$$0.4 \times 0.1 = 0.04 (4\%)$$

よって、A君が合格する確率は、

$$0.6 + 0.04 = 0.64 (64\%)$$

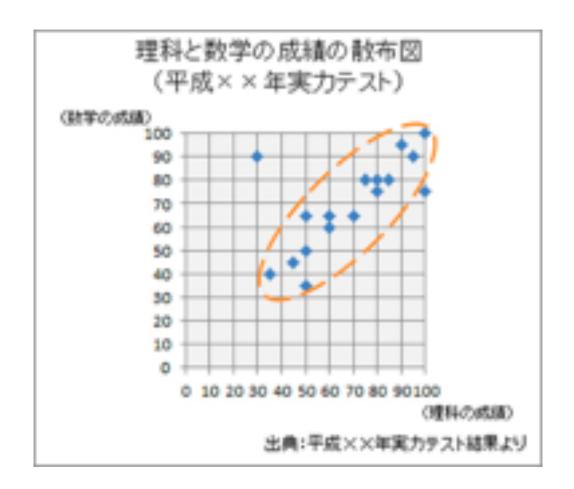
データの関係性

データの関係性を調べる

- 相関係数
 - ・関係性を数値化する
- ・クロス集計
 - ・関係性を表にまとめる

相関係数

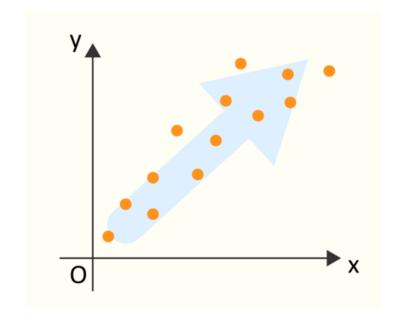
・ 関係性を数値化する



相関係数

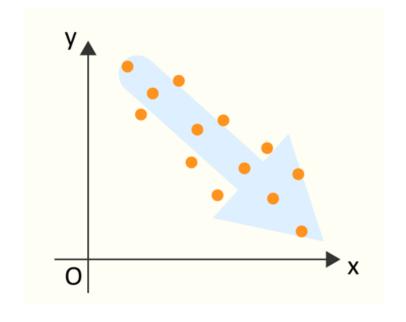
正の相関

WIN-WINの関係



負の相関

WIN-LOSEの関係



相関係数の大きさの目安

相関係数 r の値	相関
$-1 \leq r \leq -0.7$	強い負の相関
$-0.7 \leq r \leq -0.4$	負の相関
$-0.4 \leq r \leq -0.2$	弱い負の相関
$-0.2 \leq r \leq 0.2$	ほとんど相関がない
$0.2 \leq r \leq -0.4$	弱い正の相関
$0.4 \leq r \leq -0.7$	正の相関
$0.7 \leq r \leq 1$	強い正の相関

• 関係性を表にまとめる

	効果あり	効果なし	計
新薬	10	2	12
プラシボ薬	8	2	10
計	18	4	22



行集計

	効果あり	効果なし	計
新薬	83%	17%	100%
プラシボ薬	80%	20%	100%
計	82%	18%	100%

列集計

	効果あり	効果なし	計
新薬	55%	50%	54%
プラシボ薬	45%	50%	46%
計	100%	100%	100%

全体集計

	効果あり	効果なし	計
新薬	45%	9%	54%
プラシボ薬	37%	9%	46%
計	82%	18%	100%

○ 演習:どのような計算を行うと感覚が養われるか

業種	企業数	従業員数
大企業	1万1,000社	1,433万人
中小企業	380万9,000社	3,361万人

上記のような表があった場合、数字力から把握できる情報は?

- ・企業数に着目した場合
- ・従業員数に着目した場合
- ・その両方から何が言えるのか?

ここから何が

言えるのか?

まず、大企業と中小企業の区別は?

	中小企業	うち小規模事業者
業種	資本金 または 従業員	従業員
製造業・その他	3億円以下または300人以下	20人以下
卸売業	1億円以下または100人以下	5人以下
サービス業	5,000万円以下または 100人以下	5人以下
小売業	5,000万円以下または 50人以下	5人以下

○ 演習:どのような計算を行うと感覚が養われるか

業種	企業数	従業員数	従業員数/1企業
大企業	11,000	14,330,000	1,303
中小企業	3,809,000	33,610,000	9
合計数	3,820,000	47,940,000	13

業種	企業数(%)	従業員数(%)
大企業	0.3%	29.9%
中小企業	99.7%	70.1%
合計数	100%	100%

引用: https://crowdsien.com/lab/?p=824

CHEAN

演習問題: 労働力調査

- 1. 2008年と2018年における、正規/非正規雇用者比率のクロス 集計を作成せよ。
- 2. 2018年の正規/非正規雇用者における、男女比を集計せよ。

												(2) A)
		2008年	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
93	役員を除く雇用者	5175	5124	5138	5167	5161	5213	5256	5303	5391	5460	5596
\$c	正規の職員・従業員	3410	3395	3374	3355	3345	3302	3288	3317	3367	3423	3476
211	非正規の職員・従業員	1765	1727	1763	1812	1816	1910	1967	1986	2023	2036	2120
30	役員を除く雇用者	2928	2874	2865	2887	2870	2885	2898	2908	2938	2957	3008
93	正規の職員・従業員	2367	2345	2324	2315	2304	2273	2267	2272	2287	2310	2339
敷	非正規の職員・従業員	560	527	540	571	566	611	631	636	651	647	669
П	役員を除く雇用者	2248	2250	2273	2280	2291	2327	2357	2395	2453	2503	2588
女	正規の職員・従業員	1043	1050	1051	1040	1042	1029	1022	1045	1080	1114	1137
	非正規の職員・従業員	1205	1200	1223	1241	1249	1298	1335	1350	1373	1389	1451
93	役員を除く雇用者	-10	-51	14	29	-6	52	43	47	88	69	136
女	正規の職員・従業員	-39	-15	-21	-19	-10	-43	-14	29	50	56	53
21	非正規の職員・従業員	30	-38	36	49	4	94	57	19	37	13	84
100	役員を除く雇用者	-19	-54	-9	22	-17	15	13	10	30	19	51
年男	正規の職員・従業員	-41	-22	-21	-9	-11	-31	-6	5	15	23	29
被	非正規の職員・従業員	21	-33	13	31	-5	45	20	5	15	-4	22
	役員を除く雇用者	11	2	23	7	11	36	30	38	58	50	85
女	正規の職員・従業員	2	7	1	-11	2	-13	-7	23	35	34	23
\perp	非正規の職員・従業員	9	-5	23	18	8	49	37	15	23	16	62

(平成30年労働力調査)

演習問題: 労働力調査

1. 2008年と2018年における、正規/非正規雇用者比率を集計せよ。

	2008年	2018年	計
正規雇用者			
非正規雇用者			
計			

	2008年	2018年	計
正規雇用者			
非正規雇用者			
計	100%	100%	100%

演習問題: 労働力調査

2. 2018年の正規/非正規雇用者における、男女比を集計せよ。

	男性	女性	計
正規雇用者			
非正規雇用者			
計			

	男性	女性	計
正規雇用者			100%
非正規雇用者			100%
計			100%

演習問題: 労働力調査

1. 2008年と2018年における、正規/非正規雇用者比率を集計せよ。

	2008年	2018年	計
正規雇用者	3410	3476	
非正規雇用者	1765	2120	
計	5175	5596	

	2008年	2018年	計
正規雇用者	66%	62%	
非正規雇用者	34%	38%	
計	100%	100%	100%

演習問題: 労働力調査

2. 2018年の正規/非正規雇用者における、男女比を集計せよ。

	男性	女性	計
正規雇用者	2339	1137	3476
非正規雇用者	669	1451	2120
計			

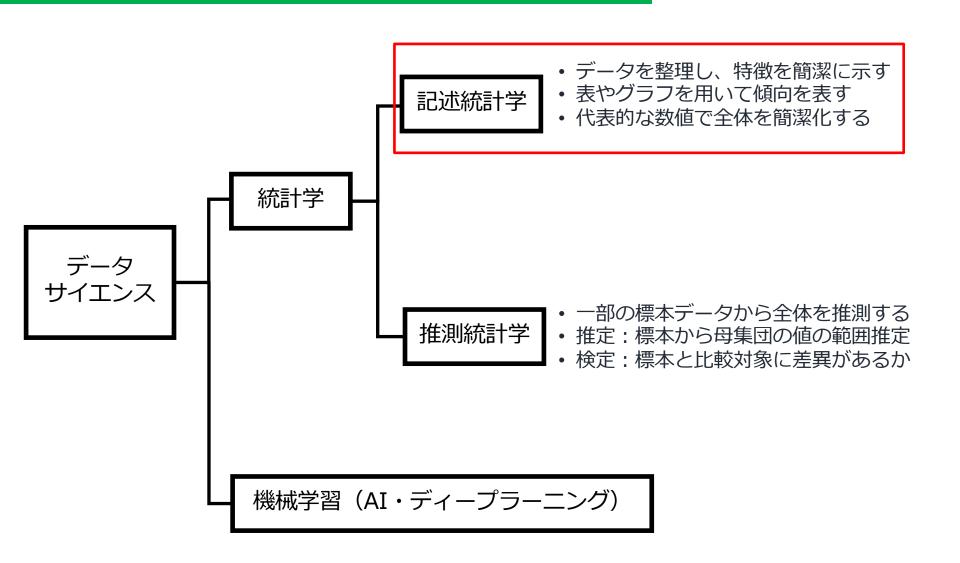
	男性	女性	計
正規雇用者	67%	33%	100%
非正規雇用者	32%	68%	100%
計			

問題演習

統計3級出題内容(2018年11月実施)

大問	小問	出題範囲	大問	小問	出題範囲
問1		データの種類	問10	[2]	データの可視化
問2		集合と確率	問11		データの集計
問3		集合と確率	問12		データの集計
問4		データの可視化	問13	[1]	相関係数/共分散
問5	[1]	データの可視化		[2]	相関係数/共分散
	[2]	データの可視化	問14		相関係数/共分散
	[3]	データの可視化	問15		相関係数/共分散
問6		データの可視化	問16	[1]	データの可視化
問7	[1]	要約統計量		[2]	データの可視化
	[2]	要約統計量		[3]	相関係数/共分散
問8	[1]	要約統計量	問17		要約統計量
	[2]	要約統計量	問18	[1]	データの可視化
問9	[1]	データの可視化		[2]	データの可視化
	[2]	データの可視化	問19	[1]	標本調査
問10	[1]	データの可視化		[2]	標本調査

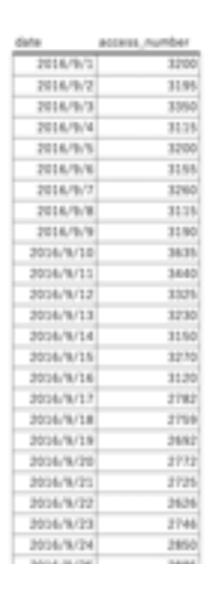
データ分析マップ



実データ

	satisfaction	evalution	projects hold Wo	sking time V	forking in of Work a	socider retirement	promoted	Department	Salary
341	9 0.36	0.47	5	134	3	9 resign	No	accounting	low
683	0.68	0.51	5	358	3	0 stay	No	technical	medium
965	0.53	0.64	2	309	3	0 stay	No	hr	medium
1220	6 0.78	0.87	4	228	5	0 resign	No	support.	low
483	6 0.92	0.56	4	170	3	0 stay	No	marketing	medium
963	7 0.98	0.82	4	179	2	0 stay	No	IT.	medium
500	5 0.69	0.81	4	264	3	0 stay	No	Sedhnical	N/W
482	3 0.66	0.85	3	266	5	0 stay	No	sales	low
903	5 0.79	0.49	4	363	3	0 stay	No	sales	high
1240	0 0.1	0.87	6	250	4	0 resign	No	sales	low
1220	6 0.87	0.9	5	254	6	0 resign	No	support.	low
656	0 0.79	0.84	4	171	3	0 stay	No	sales	low
1379	5 0.96	0.48	4	298	2	0 stay	No	sales	medium
429	4 1	0.84	3	354	3	0 stay	No	sales	medium
1290	6 0.91	0.9	5	262	3	0 stay	No	sales	medium
905	0 0.54	0.41	6	342	3	0 stay	No	groduct,mng	medium
113	8 0.87	0.88	5	262	6	0 resign	No	sales	low
686	6 0.23	0.88	5	258	6	0 stay	No	RandD	medium
1179	5 0.79	0.65	3	255	10	0 stay	No	technical	low
834	2 0.81	0.84	4	204	5	G stay	No	sales	medium
900	0.82	0.99	3	259	3	G stay	No	product_mng	
3453	6 0.93	0.89	3	255	2	1 stay	No	sales	medium
901	5 0.61	0.75	2	300	4	0 stay	No	technical	Now.
3106	2 69	0.73	2	203	4	0 stay	No	support	medium
141	5 0.55	1	6	136	3	0 resign	No	support	low

実データ

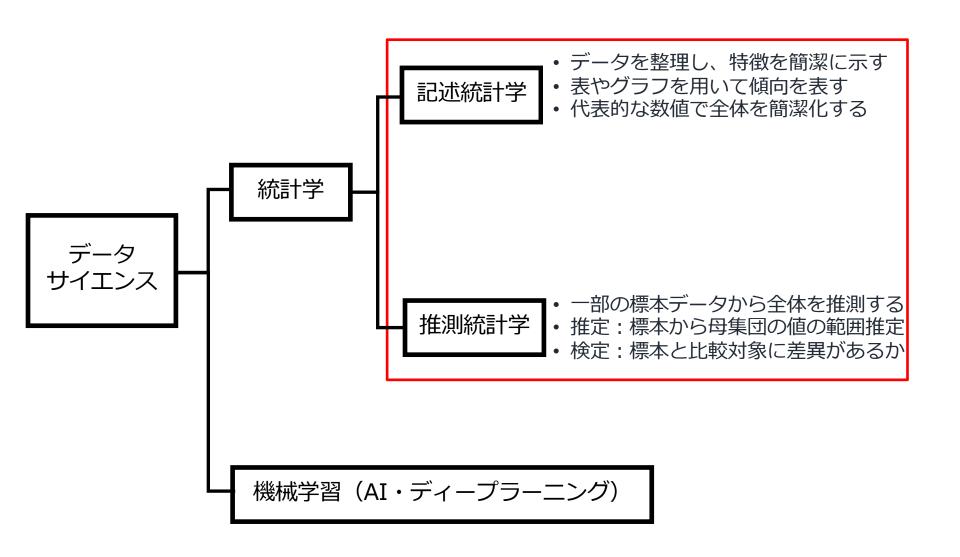


データの要約/可視化の方法は? (ピボットテーブル)

実データ

pclass	survived	sex	age	sibsp	parch
3rd	survived	female	0	1	2
3rd	died	male	0	0	2
3rd	survived	male	0	0	1
2nd	survived	male	0	1	1
3rd	survived	female	0	2	1
3rd	survived	female	0	2	1
3rd	died	male	0	1	1
2nd	survived	male	0	0	2
2nd	survived	male	0	1	1
3rd	survived	male	0	0	1
1st	survived	male	0	1	2
2nd	survived	female	0	1	2
2nd	survived	male	1	2	1
2nd	survived	female	1	1	2
2nd	survived	male	1	0	2
3rd	survived	male	1	1	2
3rd	died	male	1	5	2
3rd	survived	female	1	1	1

データ分析マップ



統計量(平均、標準偏差)を使った分析

平均と標準偏差を使って分析するとは?



統計量(平均、標準偏差)を使った分析

移動平均法



統計量(平均、標準偏差)を使った分析

異常値検出



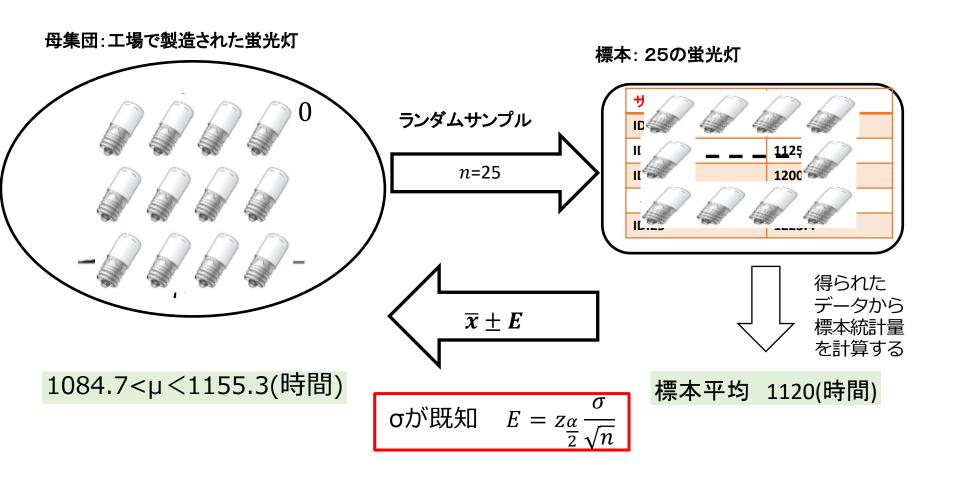
推測統計

推測統計(一部から全体を予測する)

データのばらつき(標準偏差)についてより深く理解する 確率論 ✓左右対称になった西洋の釣鐘と似た 形状の曲線(ベルカーブ) √平均値μ、標準偏差σの2つの パラメータが決まると形が決まる。 μ

区間推定の求め方

工場で製造される蛍光灯の平均寿命の95%信頼区間を求める



施策の効果測定(ABテスト)

_
_
ons

Type A

Join ABCSPORTS	
Username:	
Email:	
Password:	
I accept the Terms and Conditions	5
100% privacy. We will never spam yo	u !
Sign up +	

施策の効果測定(ABテスト)

Type A

					-	6/1
1 日平均 445	430	390	 521	560	333	250

Type B (100% privacy. We will never spam you!)

1日平均	6/30	6/29	 6/4	6/3	6/2	6/1
405			 412	462	253	159

タイプAとBのサインアップ数の間に違いがあるかのか?

データの関係性の検証

	効果あり	効果なし	計
新薬	10	2	12
プラシボ薬	8	2	10
計	18	4	22



この結果から、効果があるといえるのか?

データの関係性の検証

	効果あり	効果なし	計
新薬	10	2	12
プラシボ薬	8	2	10
計	18	4	22



新薬が標準薬(プラシボ薬)と薬効に差がない場合、 この結果はどのくらいの確率で起きることなのか?

薬効に差があるかを認めるのに必要なデータ量はいくつなのか?(サンプルサイズの設計)

データから予測を行う"重回帰分析"

重回帰分析 今までの結果から予測を行う

店舗	売上	店舗面積	アクセス数	町の規模	同業店舗数
1	231	3	294	8.2	11
2	156	2.2	232	4.1	12
3	10	0.5	149	4.3	15
4	519	5.5	260	16.1	1
5	437	4.4	567	14.1	5
6	487	4.8	237	12.7	4
7	299	3.1	451	10.1	10
8	195	2.5	433	8.4	12
9	20	1.2	212	2.1	15
10	68	0.6	502	4.7	8
11	570	5.4	378	12.3	1
12	428	4.2	577	14	7
13	?	3	400	5	3

売上はいくら見込まれるか?

データから予測を行う"重回帰分析"

概要

回帰統計					
重相関 R	0.995085404				
重決定 R2	0.990194961				
補正 R2	0.987860428				
標準調差	21.16130807				
観測数	27				

分散分析表

	自由级	流(素)	分散	観測された分散比	有意ド
回帰	5	949676.5317	189935.3063	424.1511825	2.4296E-20
75.2E	21	9403.820141	447.8009591		
合計	26	959080.3519			

	4年撤	標準調整	ŧ	P-áir	下限 95%	上限 95%
切片	-39.1953195	36.61749065	-1.07039884	0.296588834	-115.34556	36.95492102
店舗面積	20.43522388	3.91825741	5.215385754	3.61064E-05	12.28676152	28.58368623
アクセス数	-0.00073248	0.027729311	-0.02641534	0.979175474	-0.05839874	0.05693378
広告	16.966648	2.142086527	7.920617487	9.6594E-08	12.5119352	21.4213608
街の規模	15.68003698	1.973031864	7.947178793	9.15653E-08	11.5768926	19.78318136
同業者店舗数	-4.04959516	1.996639687	-2.02820528	0.055412839	-8.2018347	0.102644377