ゼロからはじめる統計学

第３回

「グラフの種類／調査法」

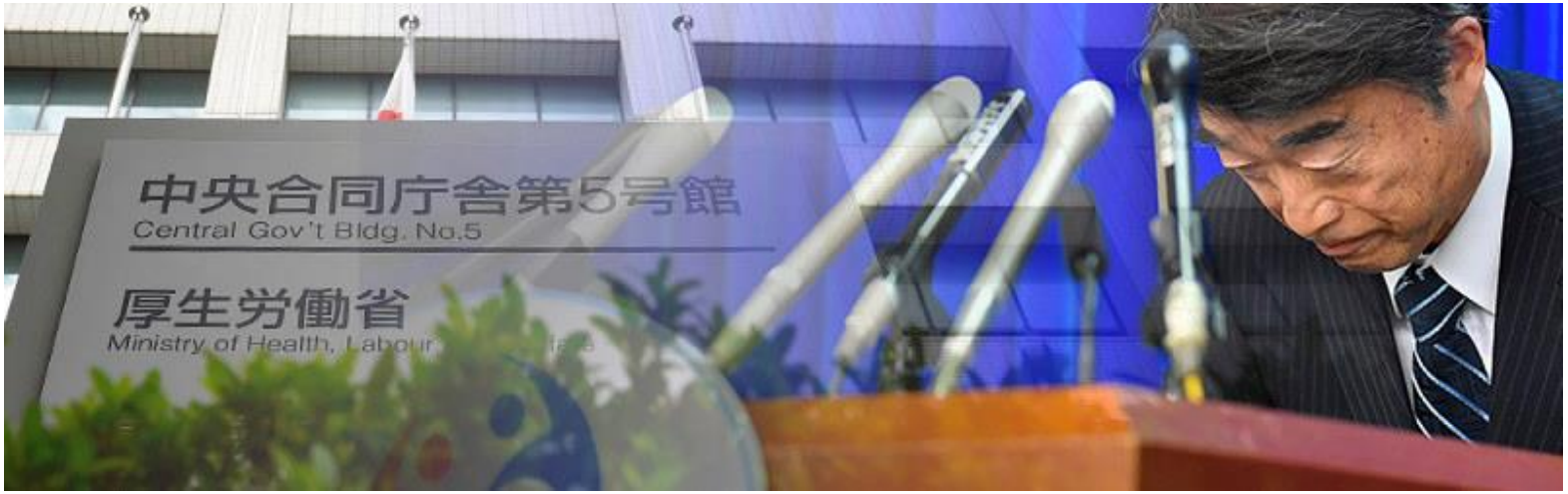
和から株式会社

サンプリング

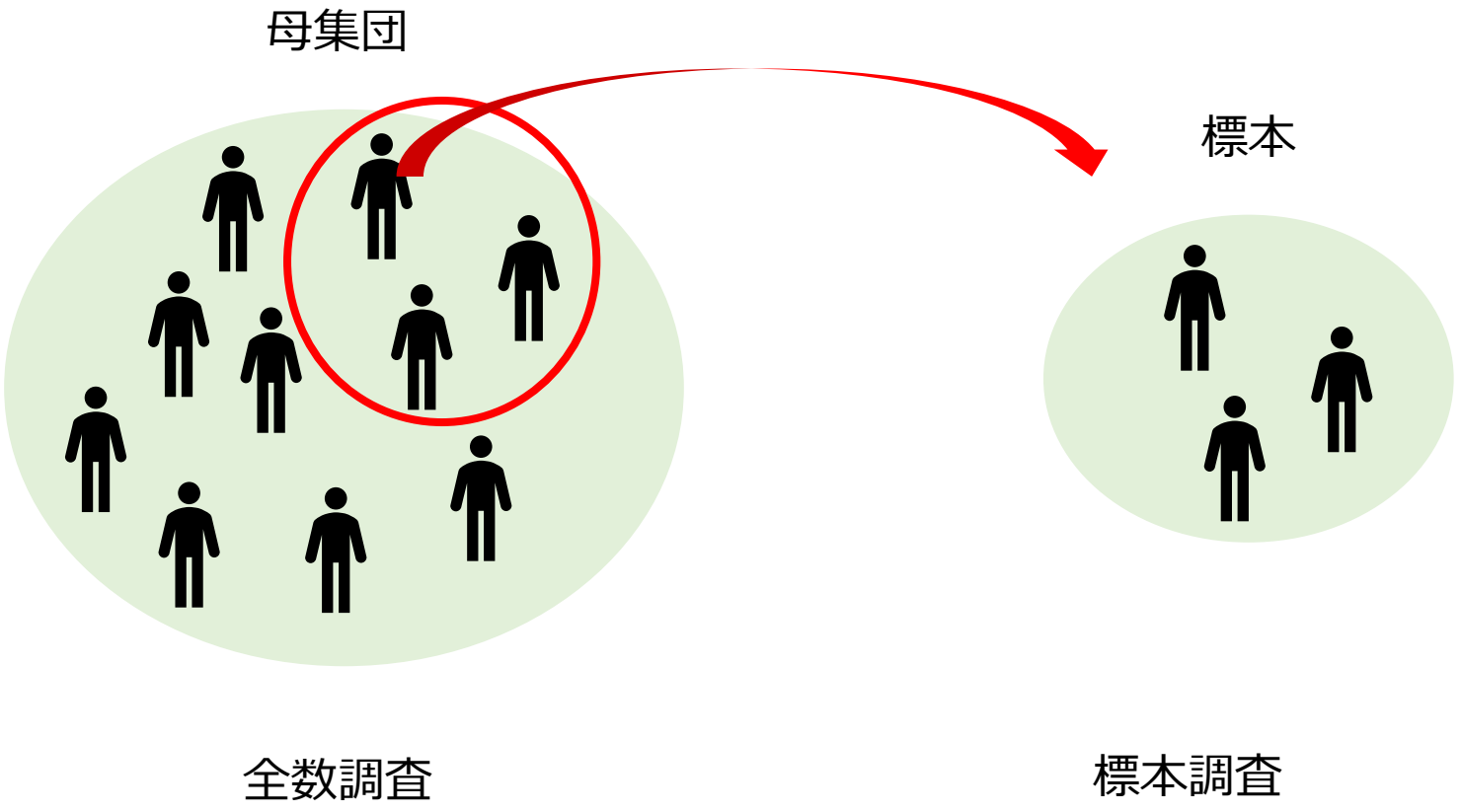
勤労統計の集計方法に関する事件

- 2018年末、国勢調査である勤労統計調査で、全数調査ではなく標本調査が行われていたことが発覚

- さらに、一部調査においてサンプリングデータの合計を全体の合計として計上



全数調査 vs. 標本調査（サンプリング）



全数調査 vs. 標本調査（サンプリング）

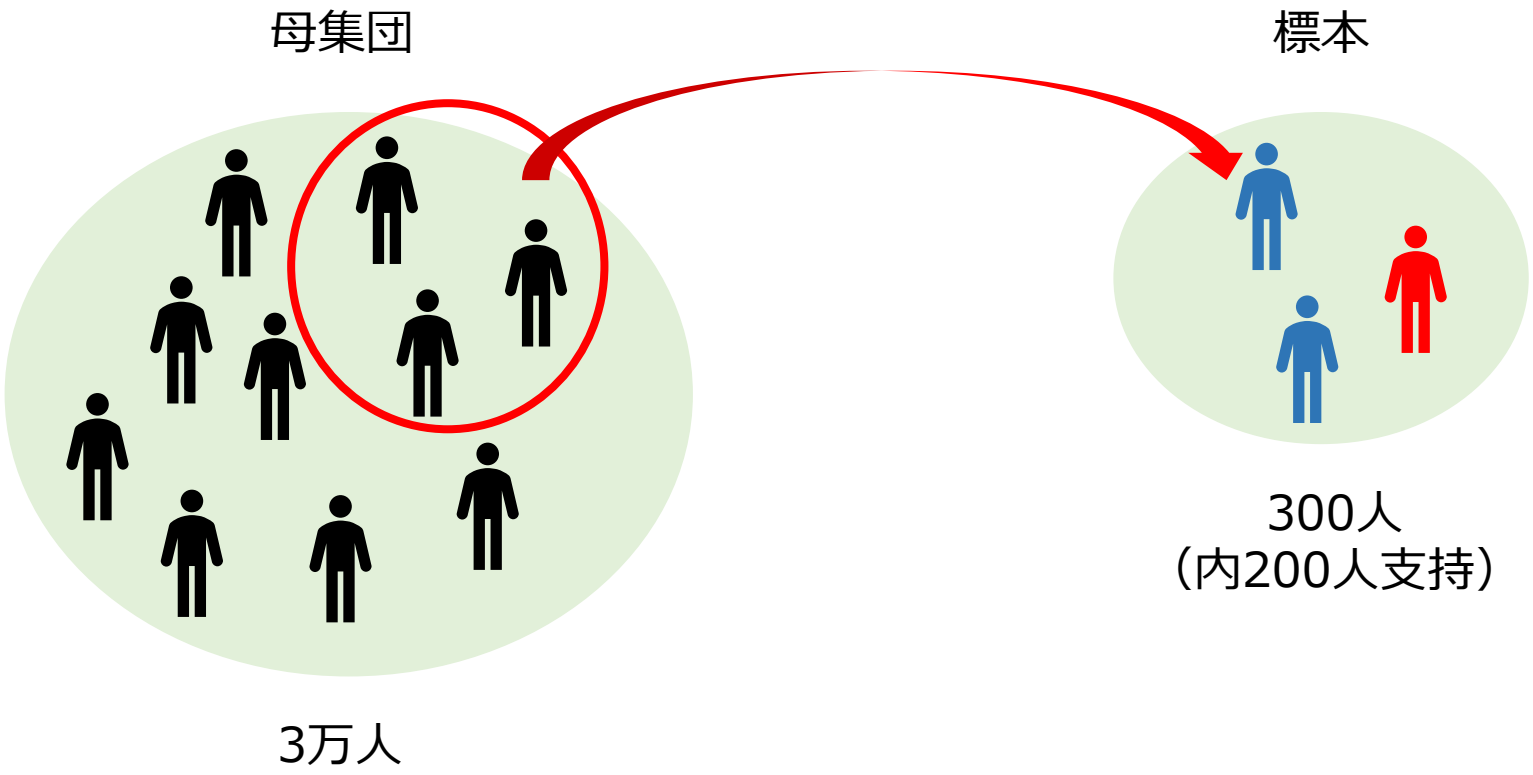
	全数調査	標本調査
内容	対象となる集団をすべて調査すること	対象となる集団を一部取り出して調査すること
メリット	正確に把握できる	時間やコストの節約になる
デメリット	コストがかかる 現実的に不可能な場合がある	標本の取り出し方によっては偏った結果が出る場合がある

国勢調査（政府統計）は、全数調査が義務付けられている

しかし…

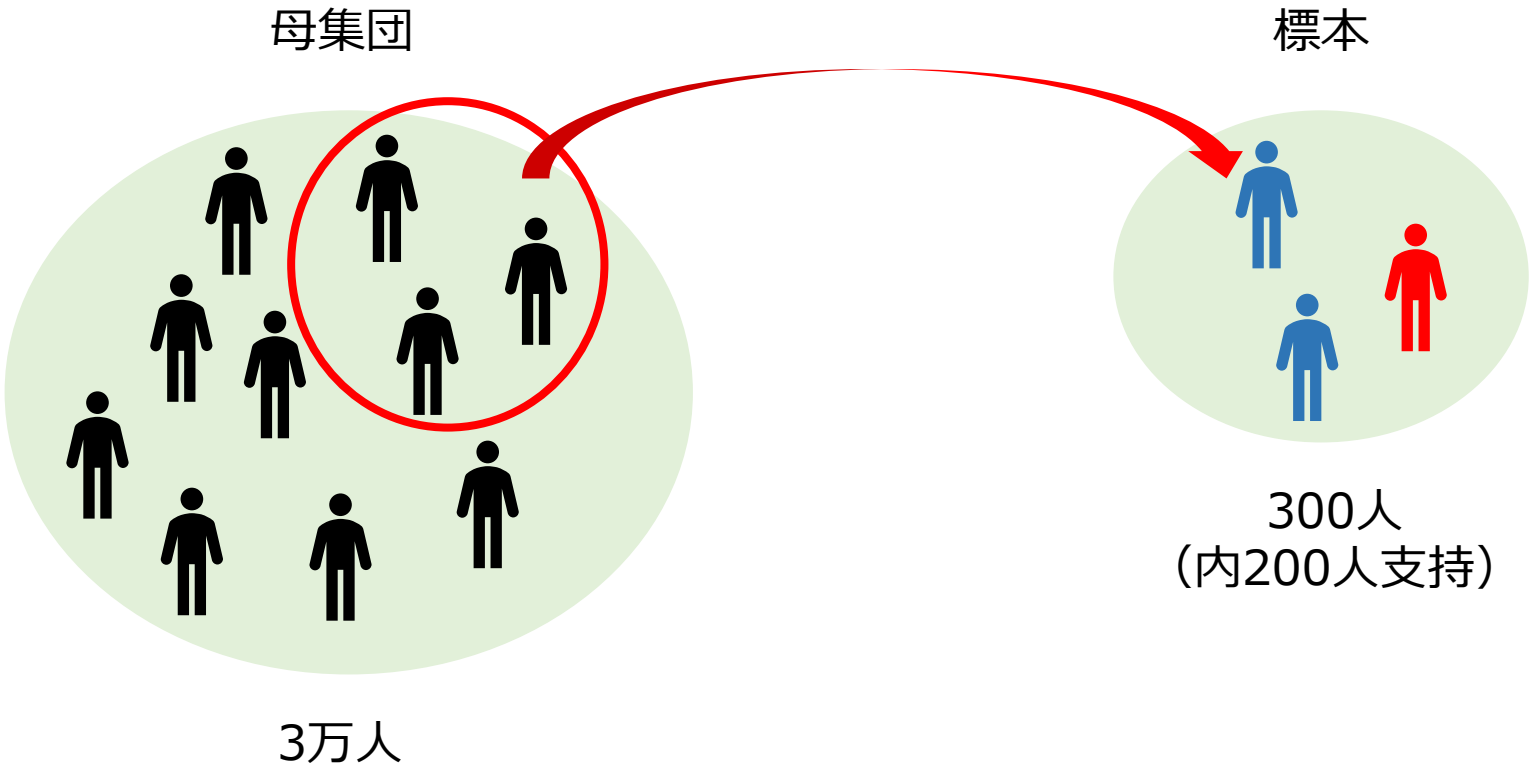
サンプリング（支持率調査）

- ある地区の3万人の有権者のうち、300人にアンケート調査を行った。するとA内閣の支持者は200人であった。この結果から、この地区でA内閣を支持する人数はどの程度いると考えられるか？



サンプリング（支持率調査）答え

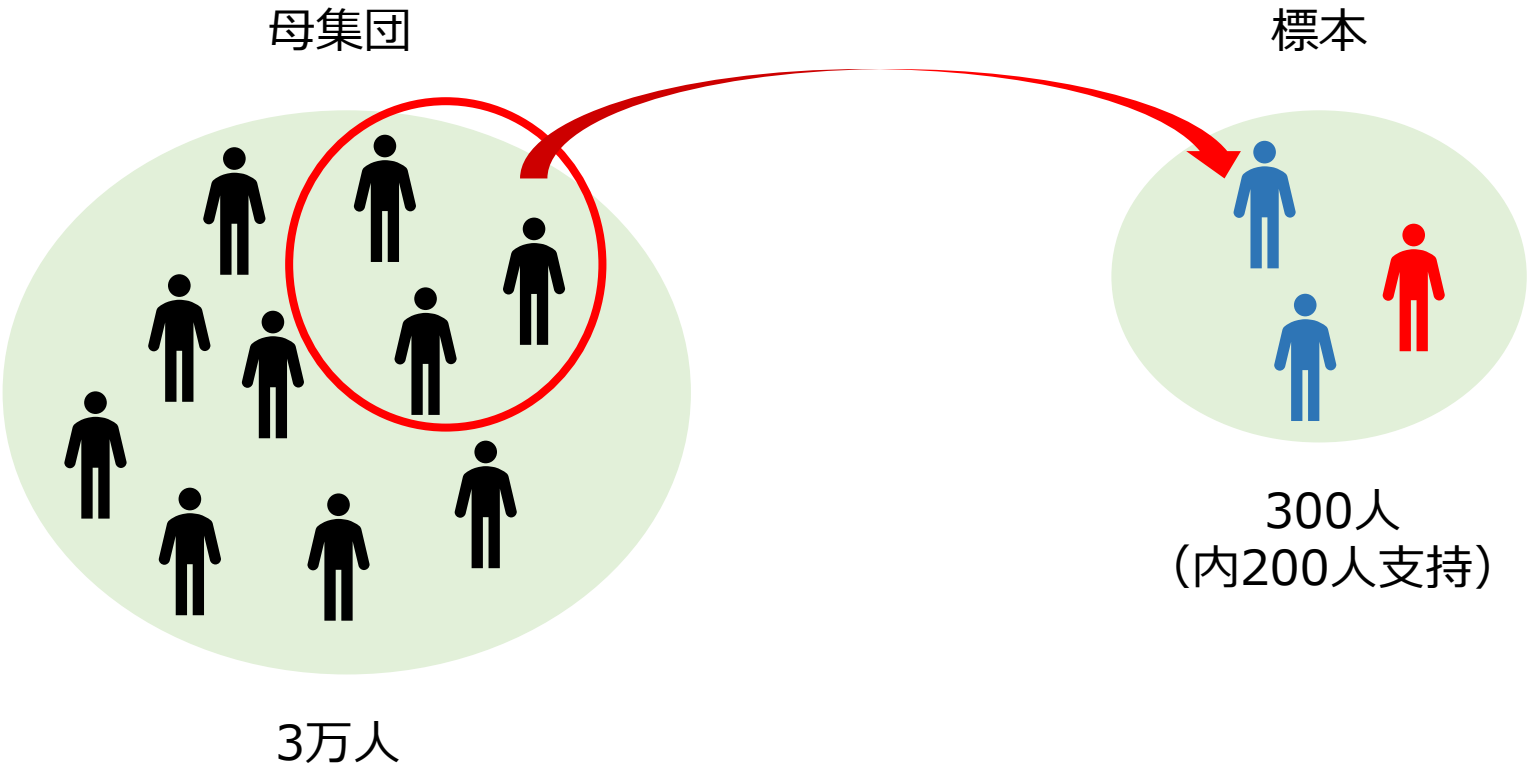
- アンケート結果から300人中200人が支持していた



- サンプリングされなかった人たちも、同じ割合で安倍内閣を支持しているのであれば、支持者の合計は、$30{,}000 \times \frac{200}{300} = 20{,}000$(人)

（忖度）サンプリング（支持率調査）答え

- アンケート結果から300人中200人が支持していた



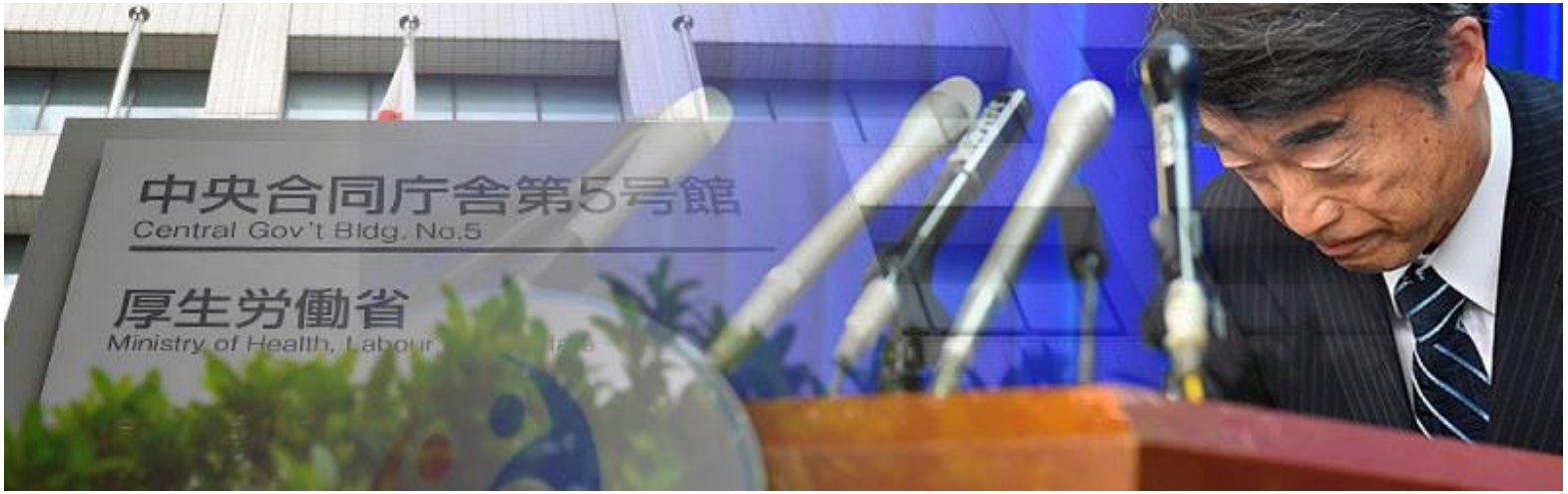
では、支持者は200人ということでよろしいですね？

残りの人たちは…?

誤った統計処理の結果

- 2004年（派遣法改正）から2018年1月まで勤労統計において、給与額が過小に集計されていた

- 勤労統計は社会保険料等の計算に使われるため、金額が上がるほど上がるはずの保険料も過小に計算された（推定600億円）



グラフ作成の手順

適切なグラフを作成するための手順

1. どのデータでグラフを作成するか決める
2. そのデータを可視化するためのグラフを選択する
3. グラフを作成する
4. 適切な凡例、タイトルなどをつける

1. どのデータでグラフを作成するか決める

ID	満足度	他者評価	プロジェクト数	労働時間（月平均）	労働時間（会社内）	Work accident	退職・在職	過去５年の昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium
56						0	在職	無	IT	medium
53						0	在職	無	technical	low
48						0	在職	無	sales	low
9335	0.79	0.49	4	163	3	0	在職	無	sales	high
12400	0.1	0.87	6	250	4	0	退職	無	sales	low
12205	0.87	0.9	5	254	6	0	退職	無	support	low

調べたい／知りたいことは何か？
伝えたいことは何か？

1. どのデータでグラフを作成するか決める

目的に応じて必要な数の項目（データ）を選ぶ

目的	必要な項目の数	例
データの多寡を知りたい	1	退職者が何人いるのか？ 全体の何%が退職しているのか？
データの分布を知りたい （分布 = ばらつき具合）	1	社員の満足度（の分布）はどうなのか？
データの関係性を知りたい	2以上	退職者とその満足度には関係があるのか？

1. どのデータでグラフを作成するか決める

ID	満足度	他者評価	プロジェクト数	労働時間（月平均）	労働時間（会社内）	Work accident	退職・在職	過去５年の昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816						0	在職	無	marketing	medium
5637						0	在職	無	IT	medium
5305						0	在職	無	technical	low
4823						0	在職	無	sales	low
9335	0.79	0.49	4	163	3	0	在職	無	sales	high
12400	0.1	0.87	6	250	4	0	退職	無	sales	low
12205	0.87	0.9	5	254	6	0	退職	無	support	low

調べたい／知りたいことは何か？
伝えたいことは何か？

2. そのデータを可視化するためのグラフを選択する

- まず、そもそもどんなグラフが使えるのか？
  - 選んだデータは量的データなのか？質的データなのか？

質的データ

✓数値でないデータ（文字型）

量的データ

✓数値データ（整数型・実数型）

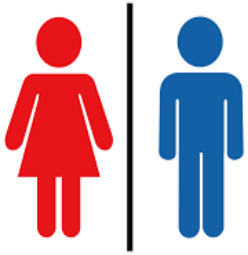
メール送信数

（１５通）

資格の有無

（有・無）

性別のデータ

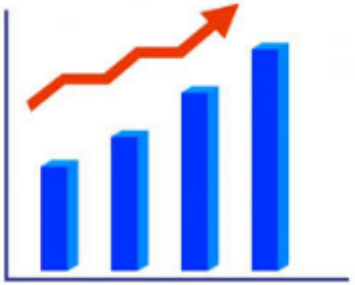


（男性・女性）

適性検査の結果

（適正・不適正）

売上



（56.7億円）

2. そのデータを可視化するためのグラフを選択する

ID	満足度	他者評価	プロジェクト数	労働時間（月平均）	労働時間（会社内）	Work accident	退職・在職	過去５年の昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium

量的データ

- 平均値
- 中央値
- 最大値
- 最小値
- 標準偏差
- ２５%、７５点
- ヒストグラム

質的データ

- 円グラフ
- クロス集計

2. そのデータを可視化するためのグラフを選択する

一変数の可視化（データ要約）

データの種類	選択可能なグラフ
質的データ	円グラフ 棒グラフ
量的データ	ヒストグラム 箱ひげ図

二変数の可視化（データの関係性）

	質的データ	量的データ
質的データ	積み上げ棒グラフ 帯グラフ	
量的データ	棒グラフ 折れ線グラフ 積み上げ棒グラフ レーダーチャート	散布図

3. グラフを作成する

ID	満足度	他者評価	プロジェクト数	労働時間（月平均）	労働時間（会社内）	Work accident	退職・在職	過去５年の昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium

① データを集計する

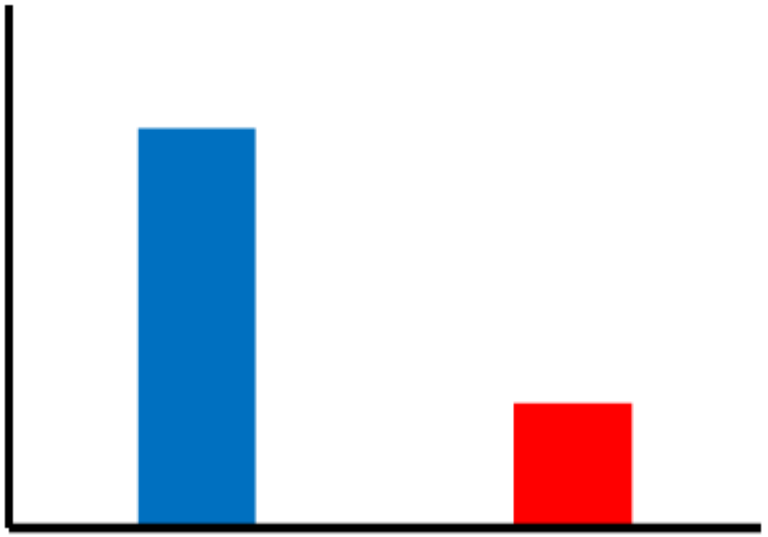
退職	在職
3571	11428

23.8%	76.2%

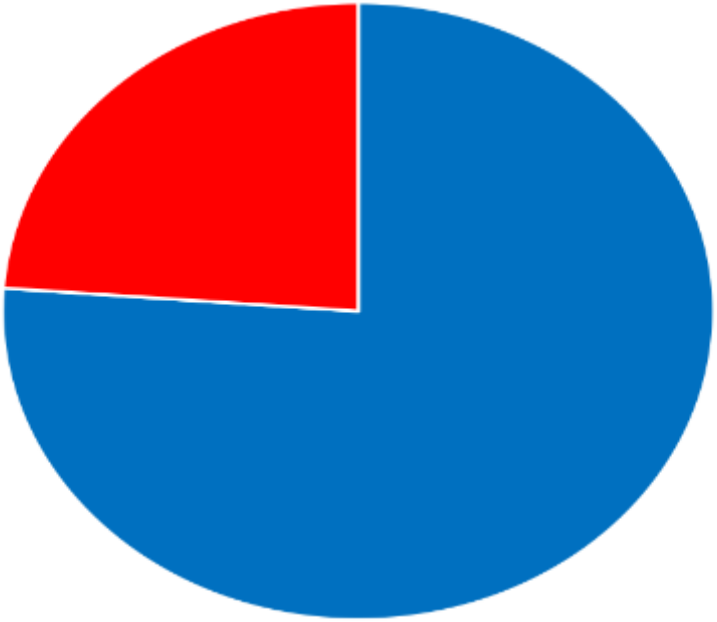
3. グラフを作成する

② 選択したグラフを作成する

棒グラフの場合

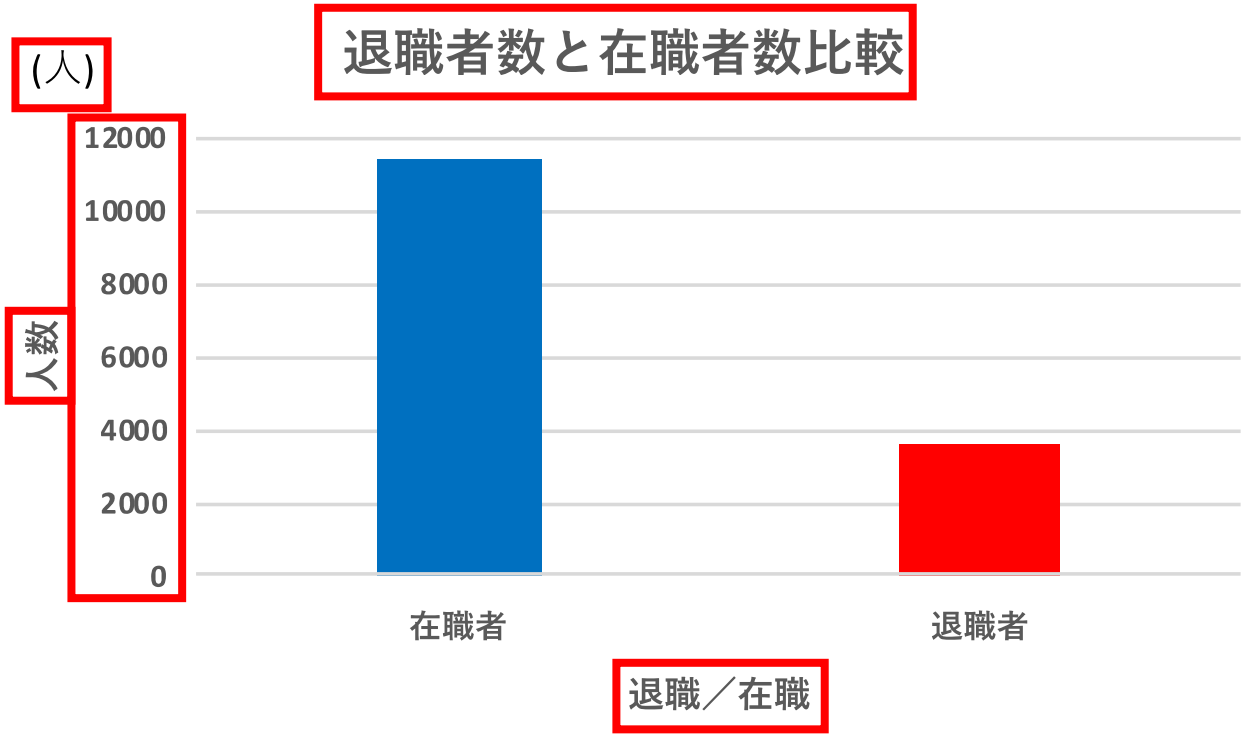


円グラフの場合



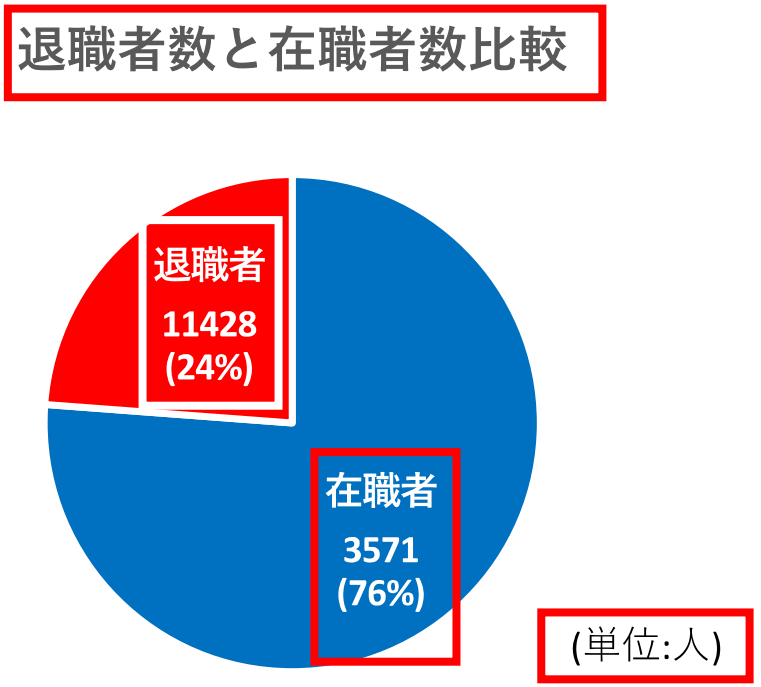
4. 適切な凡例、タイトルなどをつける

項目	目的
タイトル	何のグラフなのか
ラベル／凡例（縦、横軸とも）	何のデータを表しているのか
目盛り	どのくらいの大きさなのか
単位	何の単位なのか

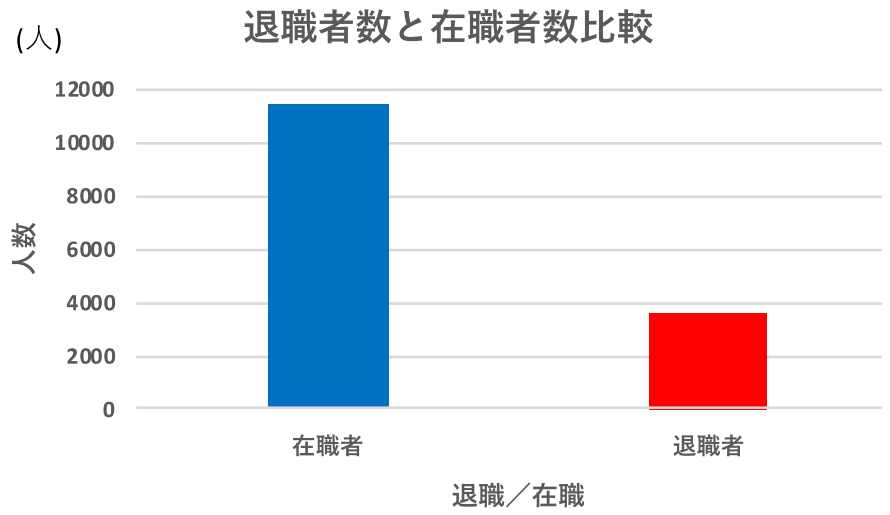


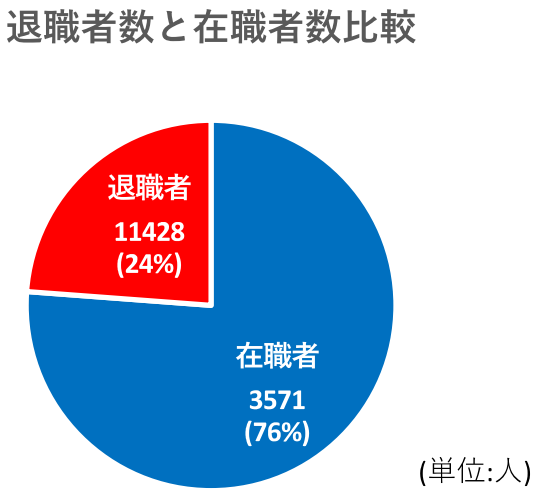
4. 適切な凡例、タイトルなどをつける

タイトル	何のグラフなのか
凡例（縦、横軸とも）	何のデータを表しているのか
目盛り	どのくらいの大きさなのか
単位	何の単位なのか



完成





棒グラフによる可視化

データの大きさが分かりやすい

円グラフによる可視化

全体に占める比率が直感的に分かりやすい

どのグラフを選択するか？

様々なグラフ

一変数の可視化（データ要約を調べる）

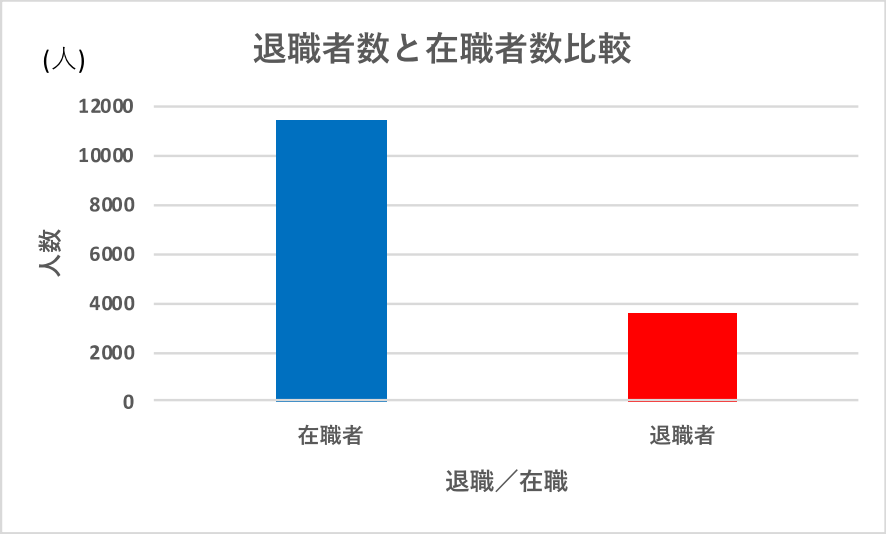
データの種類	選択可能なグラフ
質的データ	円グラフ 棒グラフ
量的データ	ヒストグラム 箱ひげ図

二変数の可視化（データの関係性を調べる）

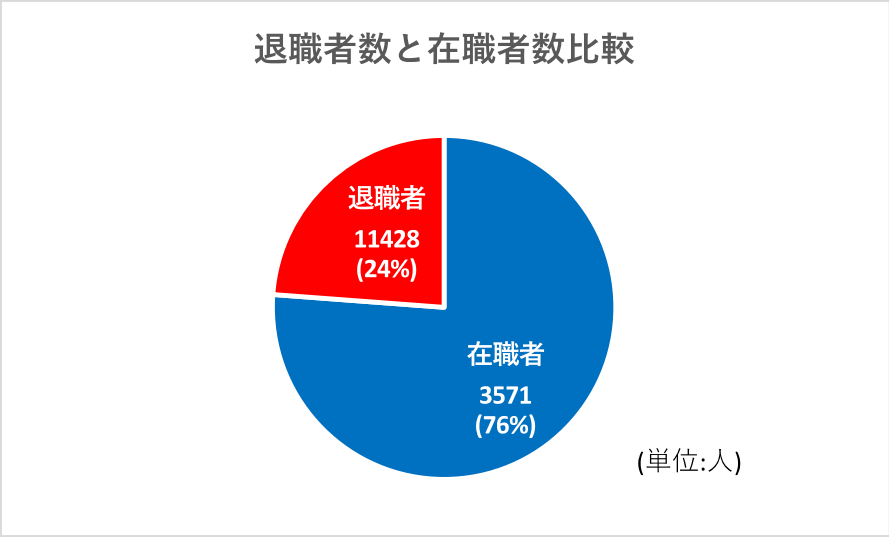
	質的データ	量的データ
質的データ	積み上げ棒グラフ 帯グラフ	
量的データ	棒グラフ 折れ線グラフ 積み上げ棒グラフ レーダーチャート	散布図

一変数の可視化（質的データ）

棒グラフ



円グラフ



データの大きさが分かりやすい

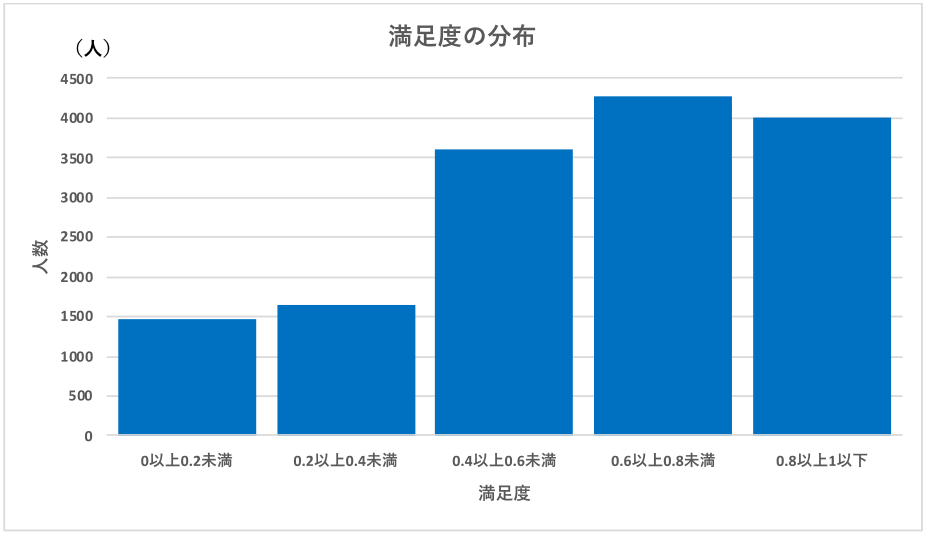
項目が多いほど、
全体に占める比率が分かりづらい

直感的に分かりやすい

目盛がなく量のイメージがつきにくい
（量が書いてない場合もある）

一変数の可視化（量的データ）

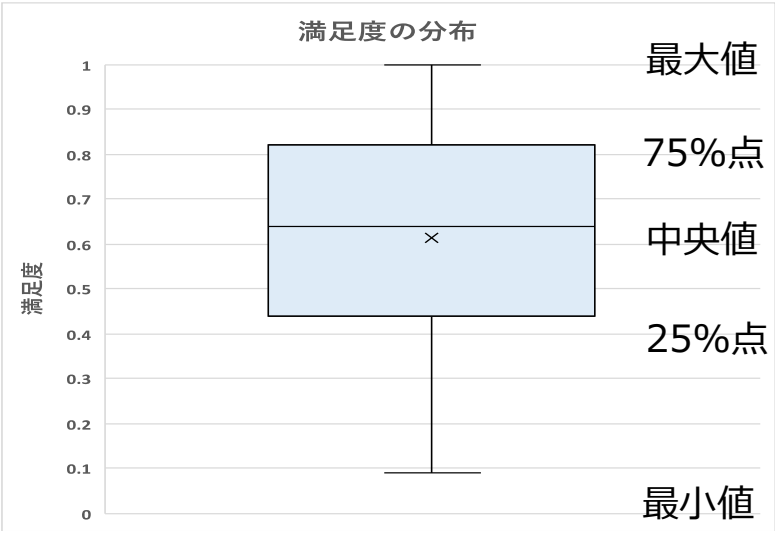
ヒストグラム



一般的に使われることが多い
データの分布（ばらつき）が見やすい

一言で要約しづらい

箱ひげ図



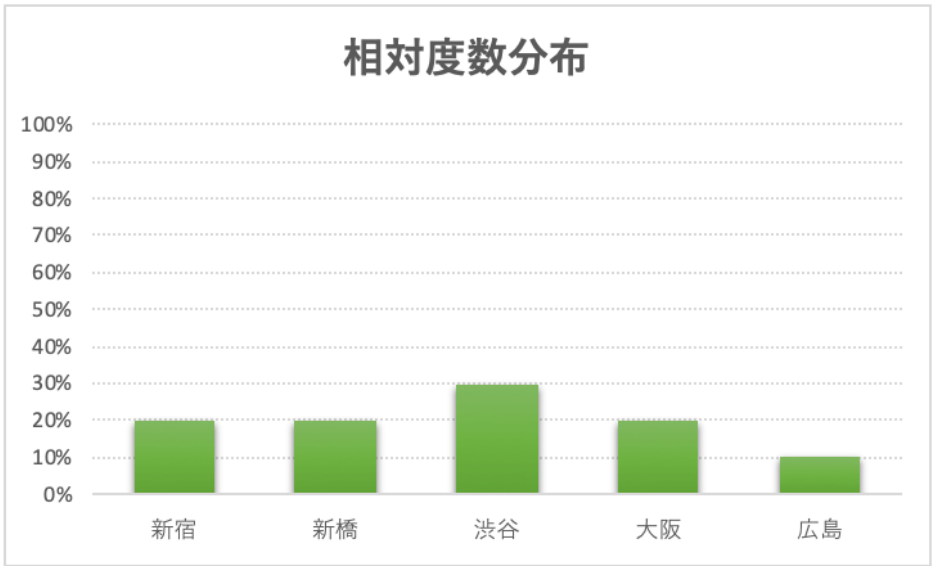
データの偏り／分布が他と比較しやすい
最大／最小値や中央値が一目でわかる

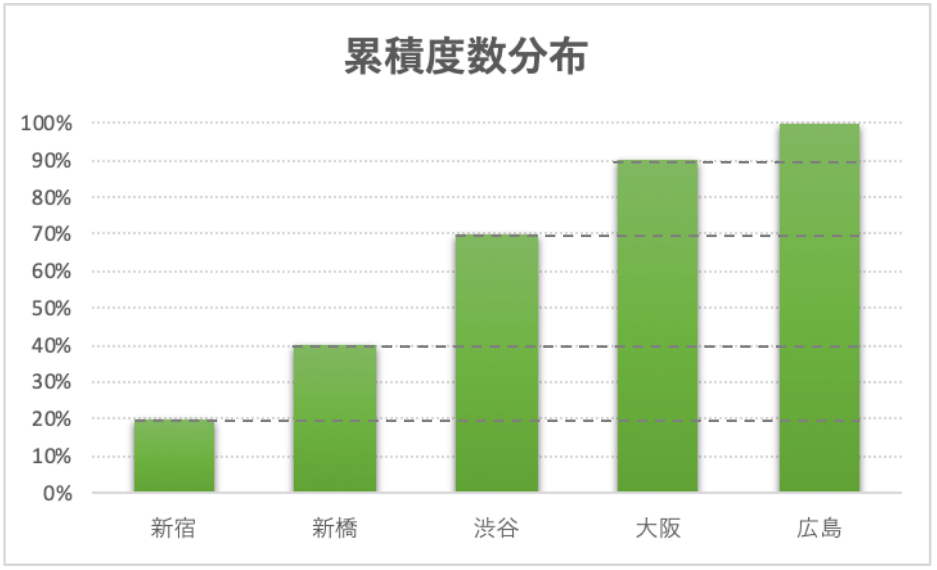
あまり知られていない

ヒストグラム(度数分布) vs. 累積度数分布

- 全体を構成するうちの何割かを示す

所在地	売上高	相対度数	累積相対度数
新宿	20	20%	20%
新橋	20	20%	40%
渋谷	30	30%	70%
大阪	20	20%	90%
広島	10	10%	100%
計	100	1	





累積度数分布の活用例：ABC分析

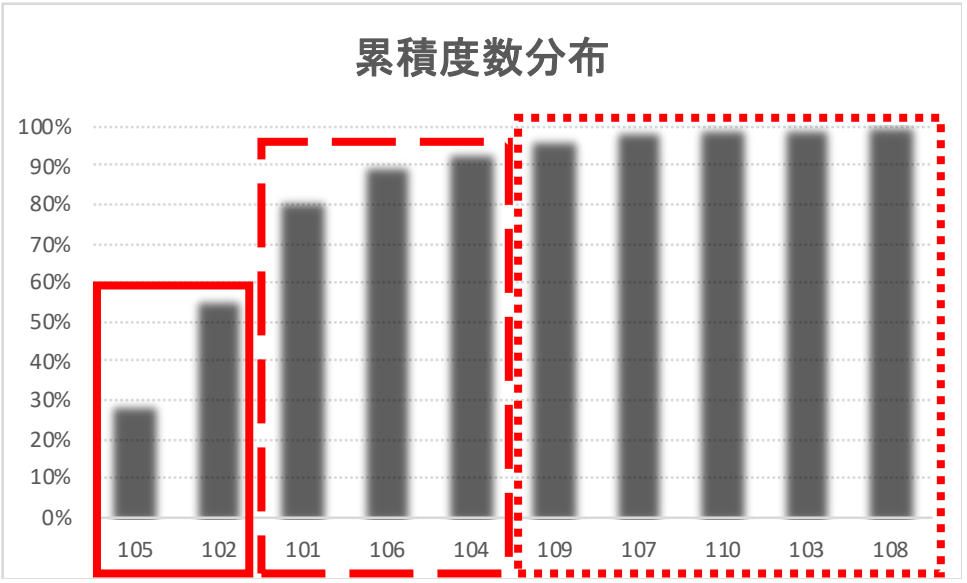
- 上位を占めるセクションに焦点を当てる分析手法

パレートの法則：
全体の数値の大部分は、全体を構成するうちの一部分の要素からなる

A…全体の80%を構成するセクション
B…全体の80%から95%を構成するセクション
C…全体の95%以下になるセクション

ABC分析

品番	売上高	累計売上高	累計売上割合	ABC分析
105	2,183,000	2,183,000	29%	A
102	2,008,380	4,191,380	56%	A
101	1,885,300	6,076,680	81%	B
106	678,300	6,754,980	90%	B
104	273,000	7,027,980	94%	B
109	218,800	7,246,780	97%	C
107	148,900	7,395,680	99%	C
110	45,700	7,441,380	99%	C
103	39,300	7,480,680	100%	C
108	21,590	7,502,270	100%	C
計	7,502,270			



A

売上全体の80%までを占める商材

B

売上全体の80%から95%までを占める商材

C

その他

ABC分析

まず、商材ごとの売上を集計する

品番	売上高	累計売上高	累計売上割合	ABC分析
101	1,885,300			
102	2,008,380			
103	39,300			
104	273,000			
105	2,183,000			
106	678,300			
107	148,900			
108	21,590			
109	218,800			
110	45,700			
計	7,502,270			

ABC分析

大きい順（降順）に並び替える

品番	売上高	累計売上高	累計売上割合	ABC分析
105	2,183,000			
102	2,008,380			
101	1,885,300			
106	678,300			
104	273,000			
109	218,800			
107	148,900			
110	45,700			
103	39,300			
108	21,590			
計	7,502,270			

ABC分析

大きい順に累計売上額を集計する

品番	売上高	累計売上高	累計売上割合	ABC分析
105	2,183,000	2,183,000		
102	2,008,380	4,191,380		
101	1,885,300	6,076,680		
106	678,300	6,754,980		
104	273,000	7,027,980		
109	218,800	7,246,780		
107	148,900	7,395,680		
110	45,700	7,441,380		
103	39,300	7,480,680		
108	21,590	7,502,270		
計	7,502,270			

ABC分析

累計売上が全体に占める割合を計算する

品番	売上高	累計売上高	累計売上割合	ABC分析
105	2,183,000	2,183,000	29%	
102	2,008,380	4,191,380	56%	
101	1,885,300	6,076,680	81%	
106	678,300	6,754,980	90%	
104	273,000	7,027,980	94%	
109	218,800	7,246,780	97%	
107	148,900	7,395,680	99%	
110	45,700	7,441,380	99%	
103	39,300	7,480,680	100%	
108	21,590	7,502,270	100%	
計	7,502,270			

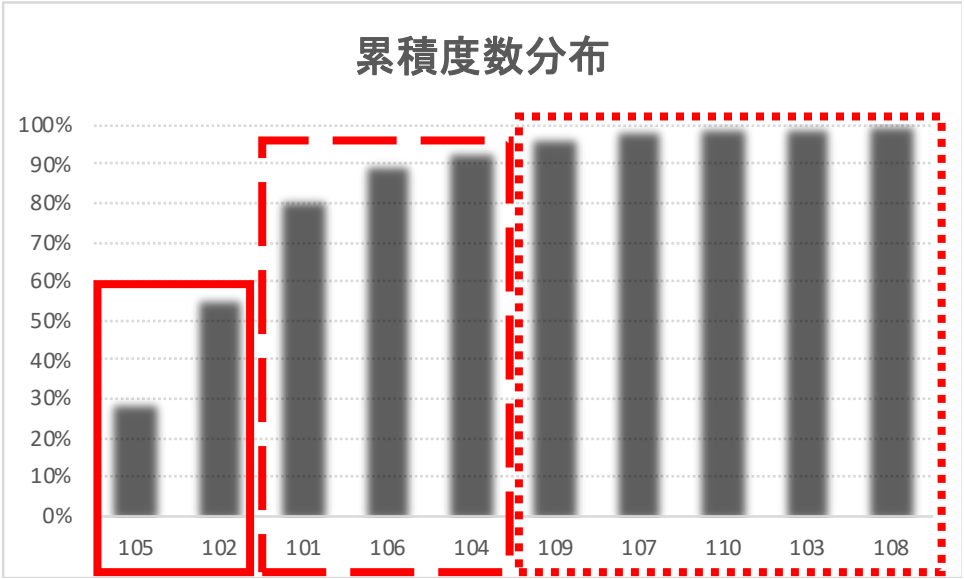
ABC分析

累積売上の80%までを占める商材をA、95%までを占めるものをB、それ以外をCとする

品番	売上高	累計売上高	累計売上割合	ABC分析
105	2,183,000	2,183,000	29%	A
102	2,008,380	4,191,380	56%	A
101	1,885,300	6,076,680	81%	B
106	678,300	6,754,980	90%	B
104	273,000	7,027,980	94%	B
109	218,800	7,246,780	97%	C
107	148,900	7,395,680	99%	C
110	45,700	7,441,380	99%	C
103	39,300	7,480,680	100%	C
108	21,590	7,502,270	100%	C
計	7,502,270			

ABC分析

品番	売上高	累計売上高	累計売上割合	ABC分析
105	2,183,000	2,183,000	29%	A
102	2,008,380	4,191,380	56%	A
101	1,885,300	6,076,680	81%	B
106	678,300	6,754,980	90%	B
104	273,000	7,027,980	94%	B
109	218,800	7,246,780	97%	C
107	148,900	7,395,680	99%	C
110	45,700	7,441,380	99%	C
103	39,300	7,480,680	100%	C
108	21,590	7,502,270	100%	C
計	7,502,270			



A

売上全体の80%までを占める商材

B

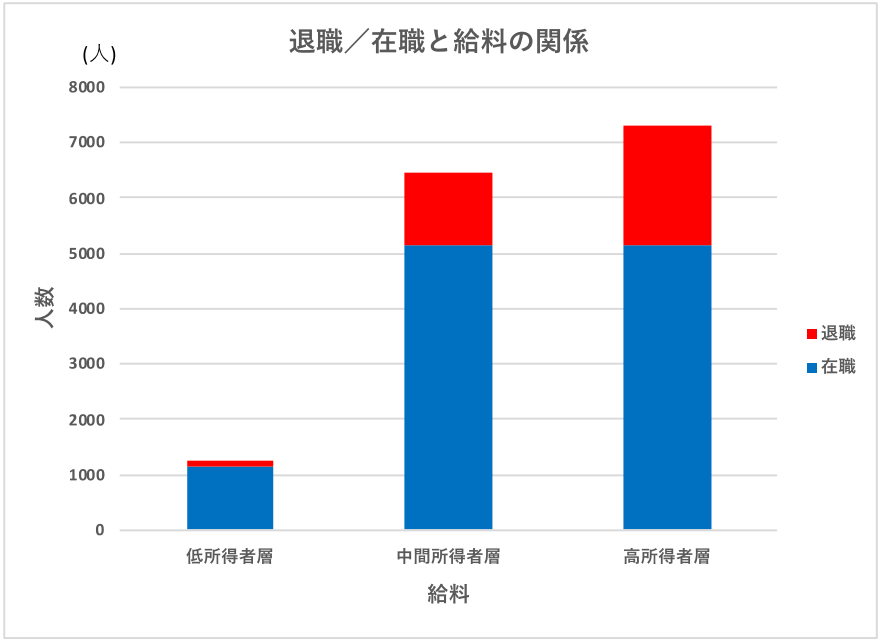
売上全体の80%から95%までを占める商材

C

その他

二変数の可視化（質的 vs.質的データ）

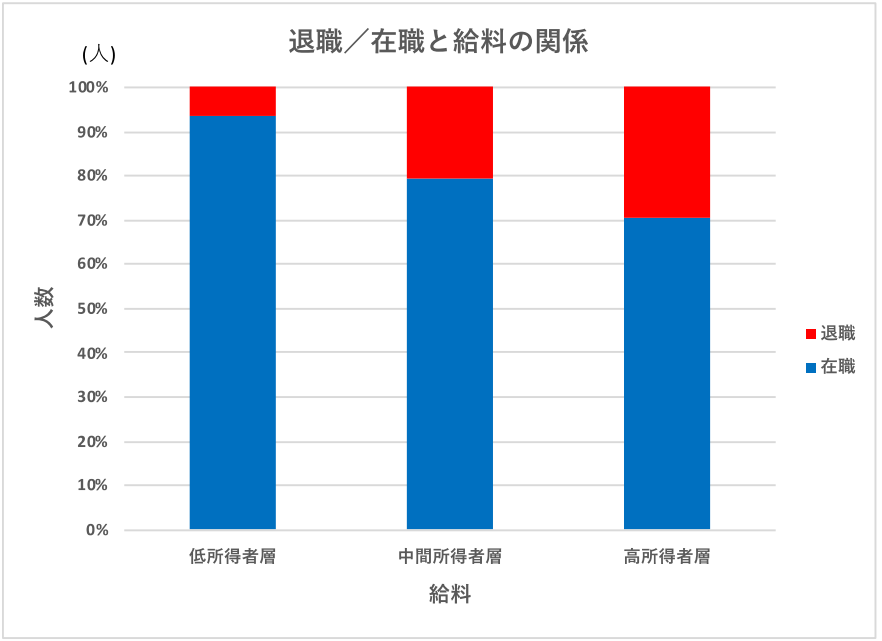
積み上げ棒グラフ



データの大きさが分かりやすい

比率が見辛い

100%積み上げ棒グラフ

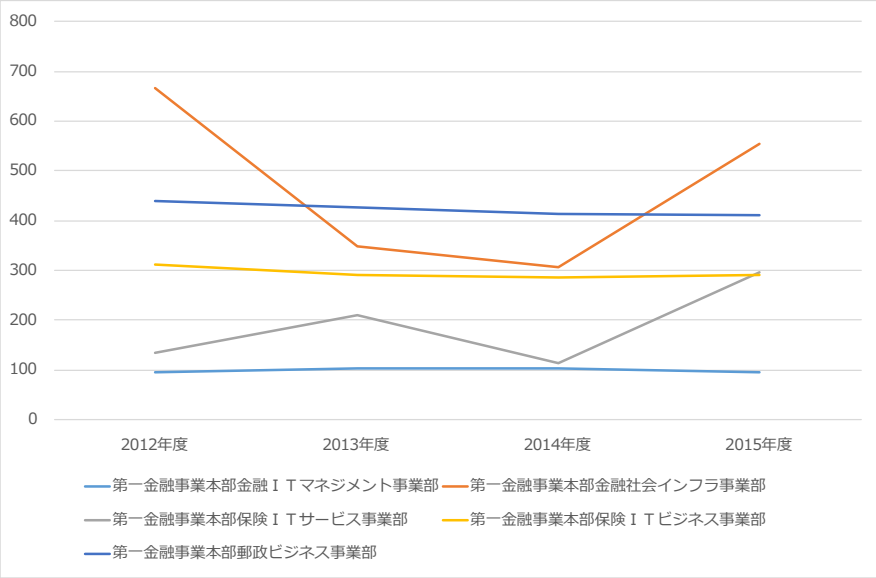


カテゴリ毎の比率を比較しやすい

それぞれの全体の人数が把握できない

二変数の可視化（質的 vs.量的データ）

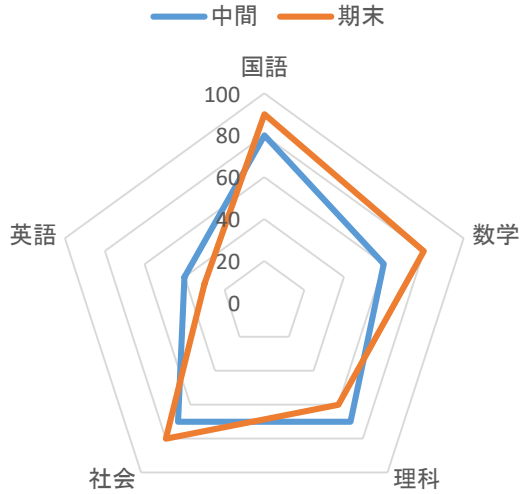
折れ線グラフ



変化が分かりやすい
複数組み合わせられる

量の大きさがやや見づらい

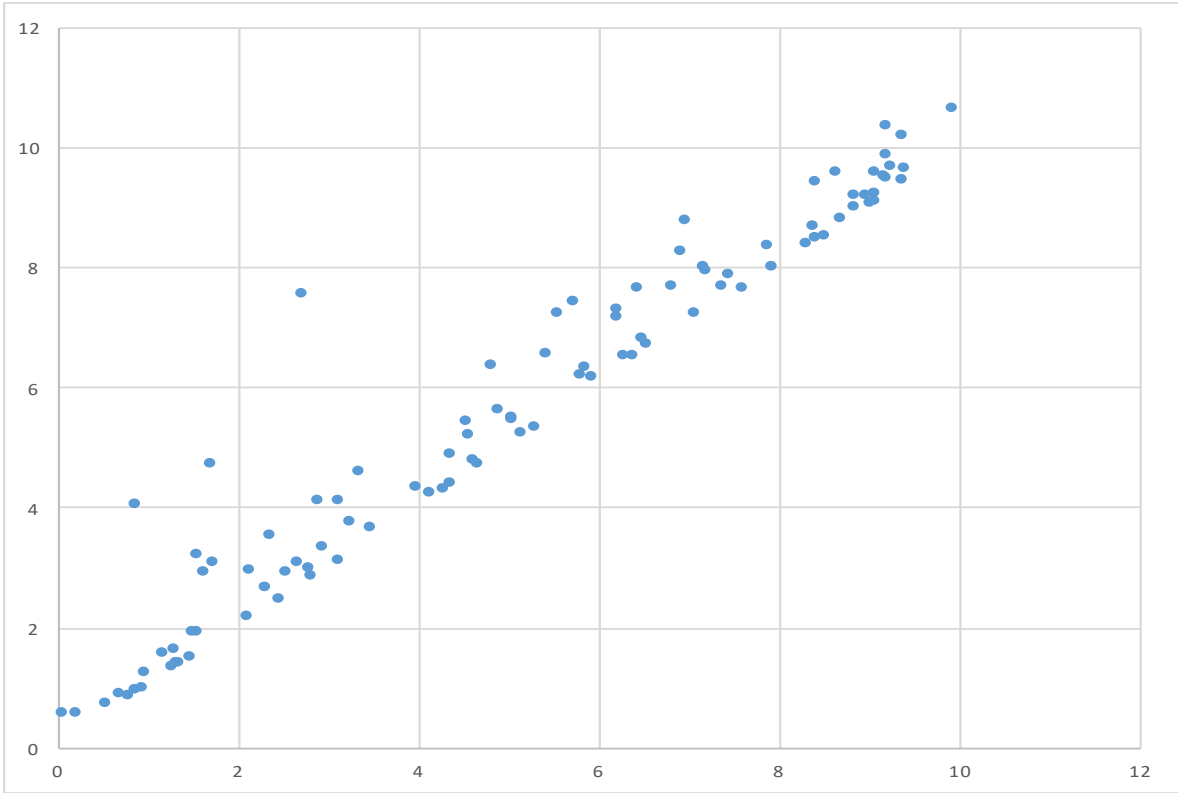
レーダーチャート



項目ごとの大小が見やすい

量の大きさがやや見づらい

二変数の可視化（量的 vs.量的データ）



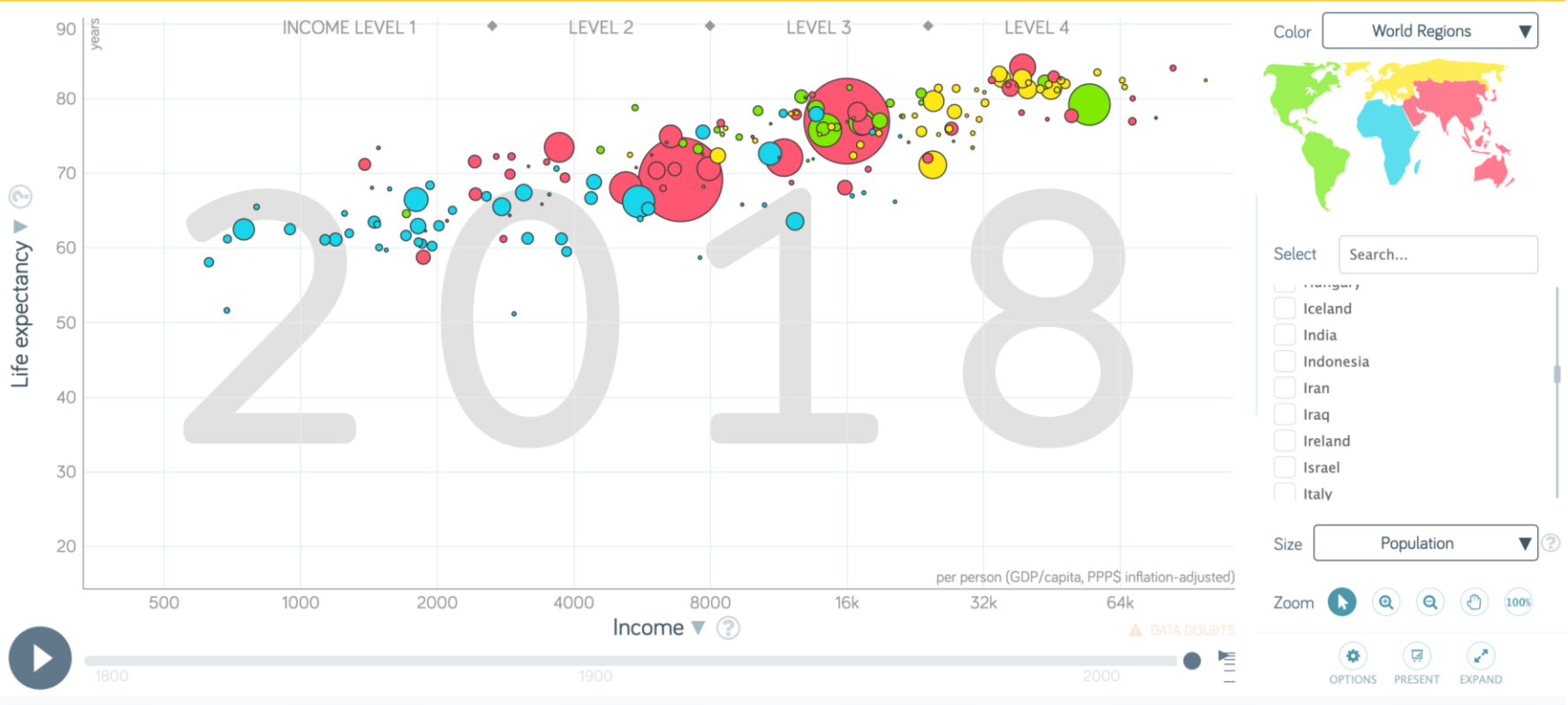
散布図

データをそのまま確認できる

自分で解釈する部分が大きい

三変数以上の可視化

バブルチャート

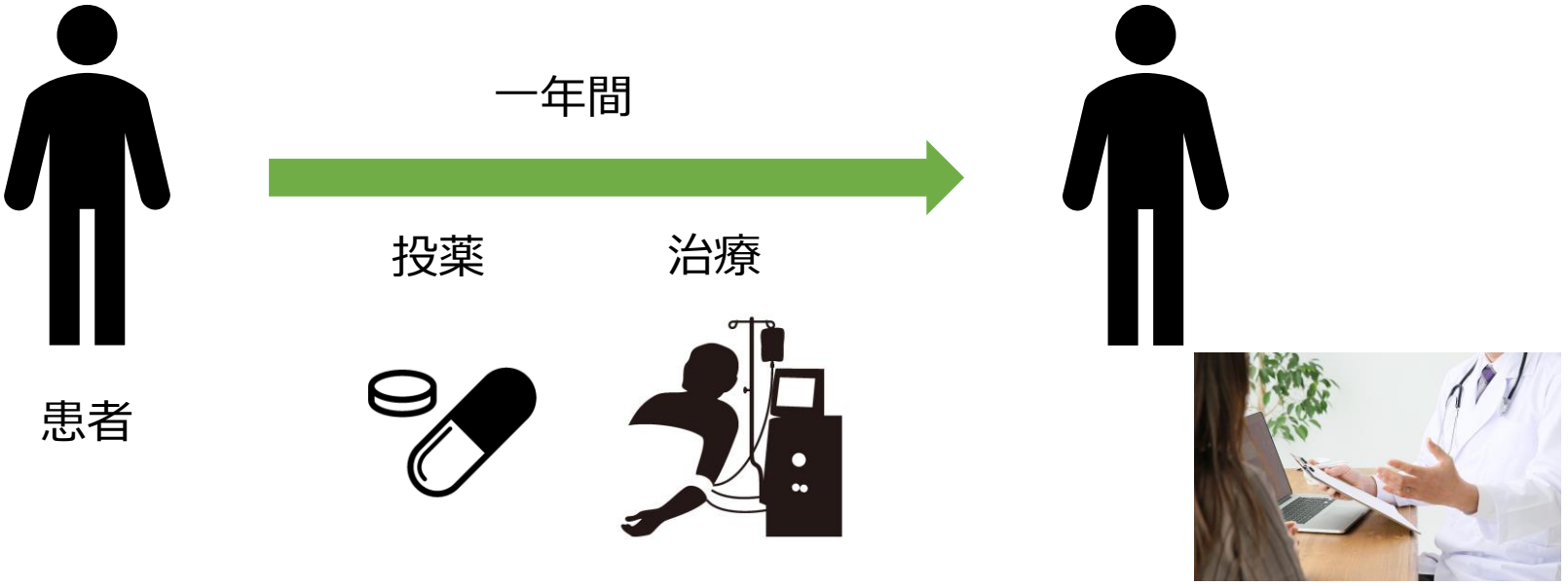


実験計画／調査法

実験計画

観察研究	実験者が被験体に介入せず、経過を観察する
実験研究	実験者が被験体に介入して、経過を観察する

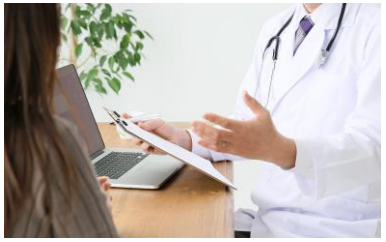
実験研究



実験計画

観察研究	実験者が被験体に介入せず、経過を観察する
実験研究	実験者が被験体に介入して、経過を観察する

観察研究



処理群と対照群（Control）

新薬の投薬試験：

	新薬投与 (処理群)	投与なし （対照群）
効果あり	10	2
効果なし	2	10

効果があるといえるのか？

処理群と対照群（Control）

新薬の投薬試験：

	新薬投与 (処理群)	プラシボ薬投与 （対照群）
効果あり	10	5
効果なし	2	7

プラシーボ効果：
薬を投与したこと自体による心理作用によって、薬理作用に基づかない効果が得られること。
（デジタル大辞泉）

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