統計学

第4，５，６回　推測統計

授業内容

- 母集団と標本
- 推定と検定の基礎
- 平均に関する推定と検定
- 比率に関する推定と検定

教科書

- P121～134（母集団と標本）
- P136～142（統計的推測の基礎）
- P143～176（区間推定）
- P178～212（仮説検定）

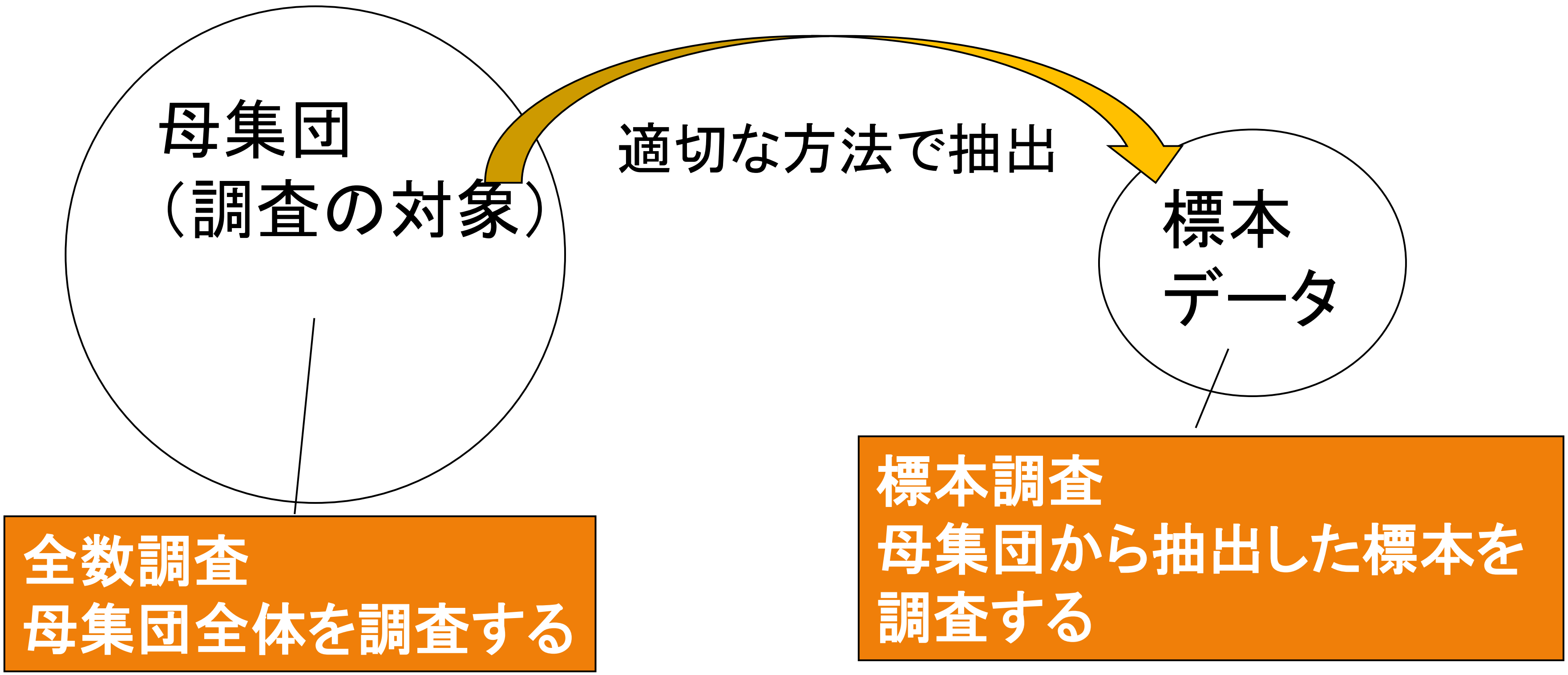
母集団と標本

統計を取る目的

　統計とは、我々の社会生活や、国や企業の経営や、科学・技術などに関する現実の問題を解決し、行動の指針を与えるような情報を提供する目的で求められるものである。

統計を作る前に、どういう統計を作るのが最もよく目的に叶うか見定める必要がある。

母集団と標本



- 標本データ：母集団から適切な方法で抽出されたデータ集合（セット）
- 大抵の統計データは「標本調査」の結果

母集団と標本

- 母集団：調査の対象になる集団
- 標本：母集団から抽出された、いくつかの個体（調査対象）からなる集団

母集団をきっちり定義する必要がある。

適切な方法で、母集団から標本を決めなければならない。

母集団と標本に関する用語集

- 要素：母集団や標本を構成する1つ1つのデータ
- 標本の大きさ：標本に含まれる要素の数
- 標本抽出：母集団から標本となるデータを選び出すこと
- 「確率」を使いたいので、無作為に抽出する
  - 大抵の標本データは「無作為」に抽出されたものだとみなされている
- 母集団分布：母集団のデータの分布状況を確率で表したもの（未知）
  - 母集団分布が分かると、母集団における統計量が分かる
  - 母集団分布が分かると、様々な確率計算ができる

全数調査と標本調査

- 全数調査：母集団を対象とした調査
- 標本調査：標本を対象にした調査

標本は、母集団の精巧なミニチュアでなければならない

↓

標本抽出法

標本抽出の例

- 世論調査を行う為に、全国の有権者から2000人を選んでアンケート調査をしたい
- どのように選べばいいか？
  - 人が集まりやすいところで適当に声をかける？
  - くじ引き？

簡単に選べて、選ばれた人が母集団（全国の有権者）の縮図になっているように標本を抽出する必要

何故「標本抽出法」が重要か

- 母集団全体を調査するのは、特に母集団が大きい場合には難しい
- 適切な標本を抽出することにより、少数の調査でも良い推測ができる
  - 1936年のアメリカ大統領選において、2つの会社が異なる方法で世論調査を行い、3000人に調査した会社は民主党のルーズベルトを、200万人に調査した会社は共和党のランドンを予想した
  - 結果：ルーズベルトが当選
  - 予想を外した会社は、電話帳や自動車登録台帳に載っている人達に質問票を送っていた（金持ちばかりが調査対象になっていた）

標本はどのくらい必要？

- なるべく大きい方が良い
  - 大きい方が、母集団の真の値に近づく可能性がある（大数の法則）
  - 正規分布の理論が使える（中心極限定理）
- 少なすぎると
  - 誤差が大きく出る
- 多すぎると
  - 調査にお金が掛かる
  - たくさん調査しても、誤差の大きさがあまり変わらない場合がある

復元抽出と非復元抽出

- 復元抽出
  - 1人取り出す→戻す→また1人取り出す
  - 同じ人が2回以上出てくる可能性
- 非復元抽出
  - 必要な人数を同時に取り出す
  - 同じ人が出てこない
- どっちが良いか？
  - 母集団が大きい場合は、どちらの抽出方法でも差がない
  - 数理統計では、復元抽出（独立性）を仮定している

標本抽出法のあれこれ

- 系統的抽出法
- ゾーンわけ抽出法
- 単純無作為抽出法
- 集落抽出法
- 確率比例抽出法
- 多段抽出法
  - 等確率抽出
  - 確率比例抽出
- 層化抽出法

（単純）無作為抽出

（単純）無作為抽出法：母集団の各要素を等しい確率で抽出し、標本とする方法

無作為抽出を行う際には、乱数さい（０～９の数字が２面ずつ書かれた正２０面体のさいころ）や乱数表が利用される

一様乱数表

０～９までの数字が
等確率で並んだ表。

〔1〕 一様乱数表 50行×40列

1	31	80	76	88	46	67	28	49	63	87	02	14	92	70	06	87	25	50	78	98
2	87	36	48	35	95	73	59	99	97	04	12	78	86	42	03	25	80	71	32	62
3	68	81	31	56	70	15	03	20	01	91	40	93	78	45	77	17	54	61	63	23
4	80	30	21	82	19	80	12	26	15	50	39	64	67	45	55	49	69	17	95	70
5	48	14	05	77	64	48	78	85	37	81	39	50	37	82	90	35	25	21	73	35
6	71	34	66	22	85	88	22	99	21	84	64	23	69	72	59	79	57	85	51	86
7	75	54	73	10	21	47	87	38	64	67	75	55	52	22	85	63	74	67	95	34
8	67	43	47	55	33	59	94	18	26	04	72	20	05	20	25	06	31	65	31	78
9	44	75	41	97	49	39	44	86	88	21	49	98	79	24	21	97	17	61	32	19
10	41	22	80	50	32	99	60	53	00	11	86	31	59	12	42	24	65	57	25	46
11	46	54	24	05	20	86	96	10	82	72	56	21	53	29	38	09	96	21	93	80
12	96	45	70	37	93	91	40	43	73	04	60	30	59	35	31	28	23	60	32	12
13	67	65	14	47	72	92	25	30	74	19	81	30	29	07	08	03	99	58	58	40
14	17	98	21	17	16	58	75	71	34	85	18	02	67	92	81	00	03	97	64	74
15	21	93	90	21	75	49	09	55	55	43	35	99	62	68	40	63	98	53	36	85
16	26	24	10	70	90	64	42	53	96	62	43	92	10	81	94	65	77	35	99	02
17	99	83	75	28	30	53	22	58	35	43	04	74	86	00	33	13	61	15	29	27
18	88	30	60	06	46	15	35	62	35	06	39	16	82	03	78	88	92	96	48	38
19	78	49	74	67	67	97	30	55	85	40	81	70	98	35	88	06	92	44	34	46
20	07	82	67	24	54	91	29	26	64	57	81	18	89	57	14	71	62	68	01	41
21	50	39	63	39	56	75	35	48	33	34	60	21	61	44	95	66	25	40	44	52
22	66	52	36	14	23	18	80	16	70	73	60	83	15	54	01	07	22	52	88	40
23	40	25	57	33	07	70	75	18	79	05	34	44	21	35	73	88	65	94	88	44
24	88	28	42	08	55	61	72	52	77	88	02	87	85	73	60	82	76	60	79	35
25	09	02	59	71	18	08	54	83	05	52	07	72	62	09	23	44	88	24	26	13
26	39	52	00	73	48	11	72	24	72	98	60	93	64	16	91	18	92	89	74	08
27	53	48	56	40	82	50	47	54	19	06	43	12	70	54	26	39	49	22	98	89
28	11	31	14	12	92	93	72	41	56	86	68	09	78	01	64	79	52	10	03	67
29	86	05	62	42	18	90	08	91	12	37	77	37	71	93	89	55	33	77	63	31
30	28	30	81	54	19	60	48	96	39	60	80	77	11	28	19	60	03	63	35	02
31	91	89	71	77	26	43	24	34	14	44	60	02	38	24	04	18	04	99	52	70
32	53	33	14	67	97	47	46	95	91	11	29	73	89	68	25	84	58	48	72	45
33	61	35	92	55	74	93	68	63	95	59	28	84	87	28	91	81	68	77	66	06
34	82	96	98	64	37	18	70	30	41	68	79	94	96	51	92	04	54	26	32	65
35	02	58	06	55	95	32	06	97	55	50	43	86	13	04	40	67	69	34	84	19
36	37	93	20	33	39	75	96	12	90	93	24	02	17	76	12	54	95	16	16	60
37	86	03	29	77	77	65	33	99	14	27	01	44	00	73	50	60	83	58	04	19
38	72	20	04	94	92	13	78	39	19	24	88	74	77	14	14	99	56	38	94	53
39	16	45	46	43	32	56	14	05	62	93	15	63	95	36	22	64	15	48	59	95
40	22	42	44	49	27	25	02	92	35	22	24	84	52	05	56	31	82	62	95	32
41	87	80	63	41	37	38	24	97	08	44	77	17	52	69	56	54	00	85	08	67
42	48	33	27	75	13	95	32	91	31	16	56	16	76	18	71	65	89	08	11	09
43	83	71	95	72	06	65	48	85	24	72	63	63	95	86	52	31	42	00	80	37
44	57	77	66	85	52	47	74	37	36	98	78	28	04	64	58	84	11	76	40	52
45	27	04	26	45	34	62	30	49	48	05	65	84	00	86	19	81	17	81	76	83
46	30	23	66	77	63	46	71	27	10	65	65	03	72	32	86	83	53	47	84	91
47	73	24	50	31	25	10	87	20	41	17	07	46	02	69	38	26	70	31	77	16
48	93	72	15	42	44	83	00	34	92	20	17	16	75	53	86	64	89	40	51	28
49	55	47	05	60	79	04	53	76	33	15	54	44	75	23	18	21	15	57	82	26
50	98	92	15	09	78	74	66	06	43	62	22	32	96	97	57	92	77	44	12	31

乱数の作成

- EXCEL：RAND()
  - 0～1までの実数をランダムに1つ生成
  - この関数を適宜調整すれば乱数を作成できる
    - INT(RAND()*(b-a+1)+a):a以上b以下の整数の乱数を作成することができる（重複あり）

実際の調査例

1　10版　2010年（平成22

世論調査　参院選

内閣支持 続落39%

消費増税「反対」48%

朝日新聞社が3、4の両日実施した全国世論調査（電話）によると、菅内閣の支持率は39%で、1週間前の6月26、27日に実施した前回調査の48%から大きく下落した。不支持率は40%（前回29%）。「いま投票するなら」として聞いた参院比例区の投票先は民主30%、自民17%、みんな6%。民主が前回の39%から大きく減らした。

＝2面に質問と回答

内閣支持率は、菅内閣発足直後の6月8、9日の調査では60%だった。発足から約1カ月で支持率がこれだけ大きく下落するのは異例だ。

支持・不支持の理由をみると、「首相が菅さんだから内閣を支持する」と答えた人が発足直後には17%いたが、今回は7%。「政策の面から内閣を支持しない」とする人が6%から16%に増えている。

消費税引き上げをめぐる菅首相の説明や対応を「評価しない」が前週の50%から63%に増え、「評価する」は30%から21%に減った。無党派層で「評価する」は13%しかいない。消費税引き上げの賛否そのものにも変化が生じ、賛成39%、反対48%。6月12、13日の調査では賛成49%、反対44%で、その後の調査では賛否が拮抗（きっこう）していたが、今回初めて反対が賛成を上回った。

反対の人のなかでみると、前回は内閣支持36%、不支持37%と並んだが、今回は支持30%、不支持50%。消費増税反対の人の離反が顕著だ。

参院選の結果、議席を伸ばしてほしい政党を挙げてもらったところ、民主26%、自民20%、みんな10%。6月12、13日の調査で民主40%、自民17%、みんな7%だったのに比べ、民主が大きく減り、みんなが増えている。

仮に参院選後に民主が連立を組むなら、相手はどこがよいかを聞いた質問では、みんな15%、自民8%、社民8%、国民新6%などだった。政党支持は民主30%（前回37%）、自民15%（同14%）など。

世論調査　質問と回答　＝1面参照

（数字は%。小数点以下は四捨五入。質問文と回答は一部省略。◆は全員への質問。◇は枝分かれ質問で該当する回答者の中での比率。〈　〉内の数字は全体に対する比率。丸カッコ内の数字は6月26、27日の前回調査の結果。連続調査として聞いた同じ質問の回答結果は、上から6月12、13日の第1回、19、20日の第2回、26、27日の第3回、7月3、4日の第4回の順に並べた。「―」はその回は実施しなかった質問）

◆菅内閣を支持しますか。支持しませんか。

	第1回	第2回	第3回	第4回
支持する	59	50	48	39
支持しない	23	27	29	40

◇それはどうしてですか。（選択肢から一つ選ぶ＝択一。上は「支持する」39%、下は「支持しない」40%の理由）

	支持する	支持しない
首相が菅さん	18〈7〉	2〈1〉
民主党中心の内閣	34〈13〉	14〈6〉
政策の面	27〈10〉	41〈16〉
実行力の面	16〈6〉	39〈16〉

◆いま、どの政党を支持していますか。

	第1回	第2回	第3回	第4回
民主党	40	32	37	30
自民党	12	13	14	15
公明党	3	4	2	4
共産党	3	2	2	3
社民党	1	1	1	2
みんなの党	2	3	3	3
国民新党	0	0	0	0
たちあがれ日本	0	0	0	0
新党日本	0	0	0	0
新党改革	0	0	0	0
幸福実現党	0	0	0	0
その他の政党	0	0	0	0
支持政党なし	30	30	29	30
答えない・分からない	9	15	12	13

◆いま行われている参議院選挙にどの程度関心がありますか。（択一）

	第1回	第2回	第3回	第4回
大いにある	37	40	34	28
ある程度ある	44	40	46	46
あまりない	16	16	16	21
全くない	3	4	4	4

◆仮にいま、参議院選挙の投票をするとしたら、比例区ではどの政党、またはどの政党の候補者に投票したいと思いますか。

	第1回	第2回	第3回	第4回
民主党	43	36	39	30
自民党	14	17	15	17
公明党	4	4	4	5
共産党	3	3	3	4
国民新党	0	0	0	0
新党改革	0	0	0	0
社民党	1	2	2	2
たちあがれ日本	0	0	1	1
みんなの党	4	5	6	6
幸福実現党	0	0	0	0
その他の政党	0	1	1	0
答えない・分からない	31	32	29	35

◆菅首相にどの程度期待していますか。（択一）

	第1回	第2回	第3回	第4回
大いに期待	17	13	―	―
ある程度期待	51	47	―	―
あまりしていない	25	32	―	―
全くしていない	7	7	―	―

◆こんどの参議院選挙の結果、民主党が単独で参議院の議席の過半数を占めた方がよいと思いますか。民主党単独では過半数を占めない方がよいと思いますか。

	第1回	第2回	第3回	第4回
過半数を占めた方がよい	44	34	36	33
過半数を占めない方がよい	44	53	50	53

◆こんどの参議院選挙の結果、どの政党に議席をのばしてほしいですか。

	第1回	第2回	第3回	第4回
民主党	40	―	―	26
自民党	17	―	―	20
公明党	4	―	―	4
共産党	3	―	―	4
国民新党	0	―	―	0
新党改革	0	―	―	1
社民党	3	―	―	3
たちあがれ日本	0	―	―	1
みんなの党	7	―	―	10
幸福実現党	0	―	―	0
その他の政党	0	―	―	0
答えない・分からない	26	―	―	31

◆今後も民主党を中心にした政権が続いた方がよいと思いますか。そうは思いませんか。

	第1回	第2回	第3回	第4回
民主党を中心にした政権が続いた方がよい	49	36	37	36
そうは思わない	36	44	43	47

◆民主党と自民党を比べてみた場合、全体として、どちらの政策がよいと思いますか。

民主党 43　自民党 25

◆民主党と自民党を比べてみた場合、全体として、どちらが政策の実行力があると思いますか。

民主党 35　自民党 36

◆消費税の引き上げに賛成ですか。反対ですか。

	第1回	第2回	第3回	第4回
賛成	49	46	46	39
反対	44	45	44	48

◆消費税の引き上げをめぐる菅首相の一連の説明や対応を、評価しますか。評価しませんか。

評価する 21（[illegible]）
評価しない 63（[illegible]）

◆菅内閣のもとで、財政のムダを減らす取り組みがどの程度進むと思いますか。（択一）

大いに進む 1（2）
ある程度進む 40（44）
あまり進まない 49（46）
全く進まない 7（5）

◆参議院選挙の後、仮に民主党がいずれかの政党と連立を組むとしたら、どの政党と組むのがよいと思いますか。

自民党8▽公明党4▽共産党1▽社民党8▽みんなの党15▽国民新党6▽たちあがれ日本1▽新党日本1▽新[illegible]幸福実現党0▽その他の政党0▽答えない・分からない55

◇

調査方法　3、4の両日、コンピューターで無作為に作成した番号に調査員が電話をかける「朝日RDD」方式で、全国の有権者を対象に調査した。世帯用と判明した番号は1885件、有効回答は1078人。回答率57%。

2010年7月5日付朝日新聞

調査方法　３、４の両日、コンピューターで無作為に作成した番号に調査員が電話をかける「朝日ＲＤＤ」方式で、全国の有権者を対象に調査した。世帯用と判明した番号は１８８５件、有効回答は１０７８人。回答率57％。

RDD調査：

電話番号をランダムに作成し、作成された電話番号に電話を掛ける

標本抽出の実際

- 実際には：無作為（ランダム）に調査対象を選ぶのは難しい（制約がある）
  - 医学：きちんと調整を行えば、ランダム化比較実験とケースコントロール研究（調査対象をランダムに決定していない）との差は出ないことが、これまでの論文を検討することによって分かっている
  - 属性が似ているもの同士を比較する

特に、比較をするときに

- 条件がフェア（公平）かどうか、が重要
- 例：暴力的なテレビゲームばっかりしている子は暴力的な傾向がある？
  - ゲームからは影響を受けない「暴力性」といった原因があり、「暴力性」が高いほど暴力的なゲームを好み、また行動も暴力的になるのかも
  - 同じゲームに対して目くじらを立てる親とおおらかな親とで違うのかも
- 普通にアンケート調査を行っただけでは、そこまで考えるのは無理

解決方法案

- 調査対象の属性を調べ、似た属性の人々について比較する
  - 「関連しそうな条件」を考え得る限り継続的に追跡調査し、統計学的な手法を用いて、少なくとも測定された条件については「フェアな比較」を行う
  - そもそものデータの取り方の時点で、「フェアに条件をそろえる」（ランダムに割り付ける）
  - 属性を調べ、似たもの同士で比較する（傾向スコア）

母集団分布

- 母集団の確率分布のこと
- 母集団分布が分かると、様々な確率計算ができる
- 母集団分布には、母集団における統計量が含まれている
  - 母平均
  - 母分散
  - 母比率
- 母集団分布は、母集団のデータ全部が分からないと分からない（未知）
  - 大抵は「正規分布」を仮定する（モデル）

母集団分布に含まれているもの

- 母平均
- 母分散・母標準偏差
- 母比率
- 母中央値、母最頻値、母最大値…

標本から
これらの値を
推測する

母数（parameter）：母集団を特徴付ける値

統計量

- 標本から算出される量
  - 標本平均
  - 標本分散
  - 標本比率

母平均と標本平均

- 母平均：母集団データから計算される平均
- 標本平均：標本データから計算される平均
- 標本平均から母平均を推定・検定する

$$\text{母平均}\ \mu : \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

$$\text{標本平均}\ \bar{x} : \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

標本平均の性質

- 見た目（計算公式）は、母平均と同じ
- 標本データから計算される
  - 標本の取り方によって違う結果が出る
  - でも、その期待値（平均）は「母平均」になるだろうと予測される

母分散と標本分散

- 母分散：母集団データから計算される分散
- 標本分散：標本データから計算される分散
- 標本分散から母分散を推定・検定する
  - 母標準偏差と標本標準偏差でも事情は同じ

$$\text{母分散 } \sigma^2 : \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 = \frac{(x_1-\mu)^2+\cdots+(x_N-\mu)^2}{N}$$

$$\text{標本分散 } s^2 : \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{(x_1-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1}$$

標本分散の性質

- 「母平均」ではなく、「標本平均」からの偏差平方和を計算している
- 「データ数」ではなく、「データ数－１」で割っている
  - 不偏分散

何故データ数−１で割るか あるいは自由度の話

- 分散の計算：先に平均を計算する必要
- 「平均」に「固定」される

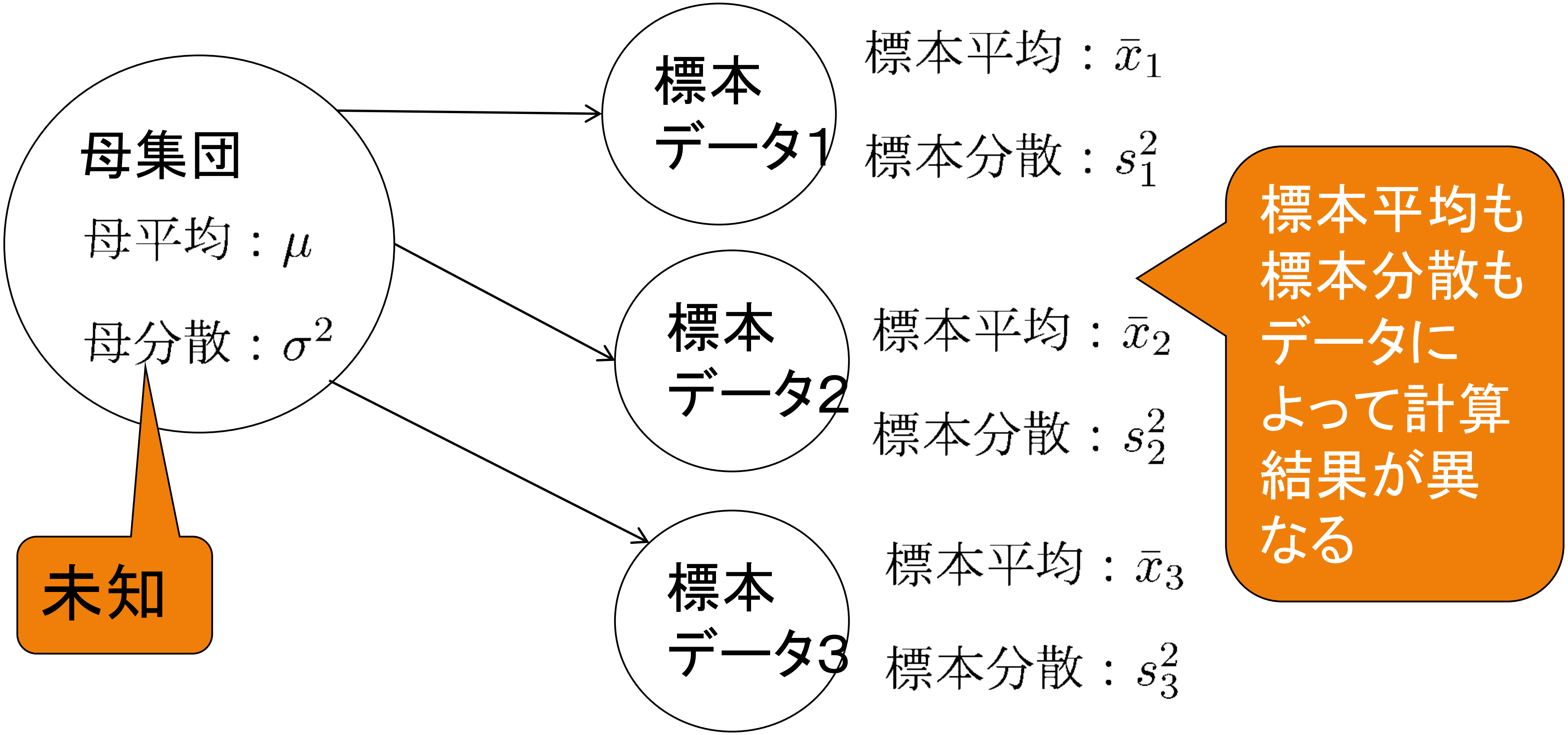
例：データが２つ（x_1, x_2）の場合：

平均が $\bar{x} = 3$、$x_1 = 4$ ならば、x_2 は計算で求めることができる。

平均が分かっているので、
データが1つ分かれば
もう一個も分かる（固定される）

自由度
（固定されていない
データの個数）

母集団と標本、標本平均、標本分散



標本分布

- 母集団のデータ：母集団分布に従っている
- 母集団から標本を抽出して平均を計算することを何度も繰り返すと、（標本データにも変動があるから）平均にも確率的な変動が見られる（標本変動）
  - 数理統計学では、データ＝確率変数とみなしている
  - 標本平均は、確率変数の和なので確率変数、すなわち、標本平均の期待値（平均）や分散を（確率として）計算することができる
- この変動を記述する→標本分布
- 標本分布をもとに、母集団の性質について考えていく
  - 母集団分布は未知（たいていは正規分布を仮定）

標本分布２

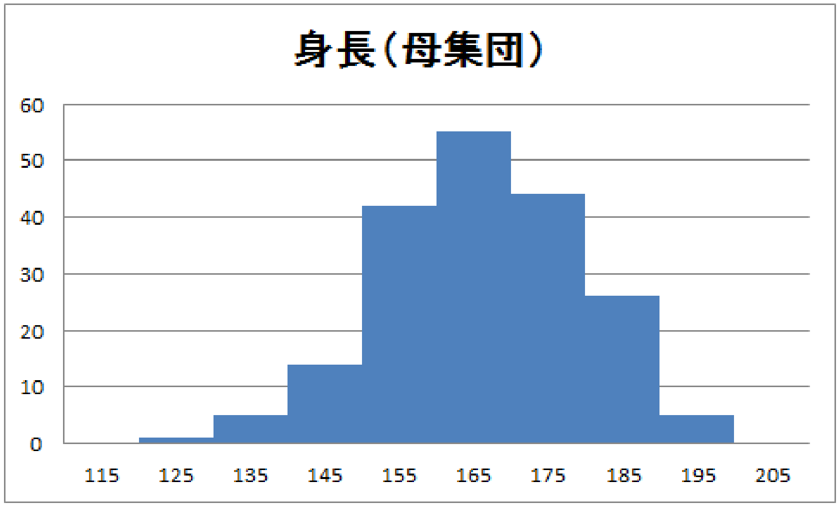
- 例えば、標本平均について、何度も標本を取って標本平均を計算することを繰り返せば、１変量のデータを得ることができる
- このデータについて、度数分布表（相対度数曲線）をつくることができる
- 平均や分散（標本変動）も計算できる

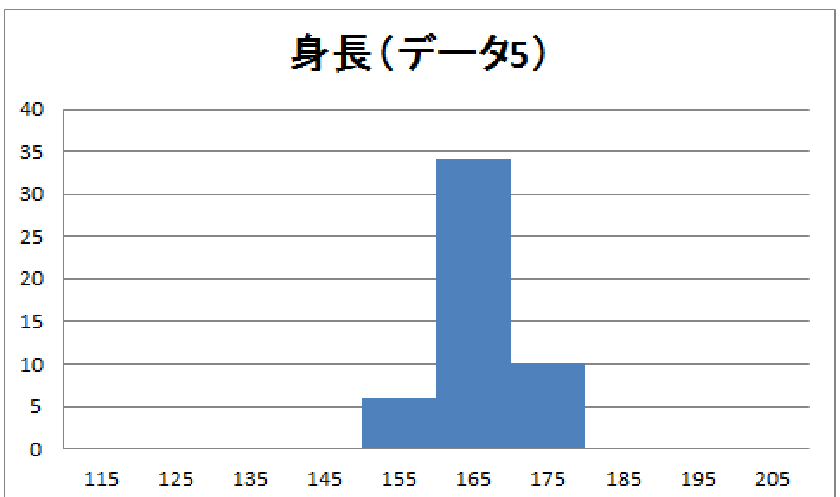
標本分布：統計量の確率分布

標本平均の平均と分散の性質

- 標本平均の平均は、理論的には母平均と同じ
- 標本平均の分散は、母分散の1/n

標本平均の平均と分散の性質例





- 上 : 192人の身長
- 下 : 上のデータを母集団として5つのデータを無作為抽出し、平均を計算することを100回繰り返したデータ
- 標本平均の平均は母集団の平均と同じところにある
- 分散は、母集団の分散よりも小さくなっている

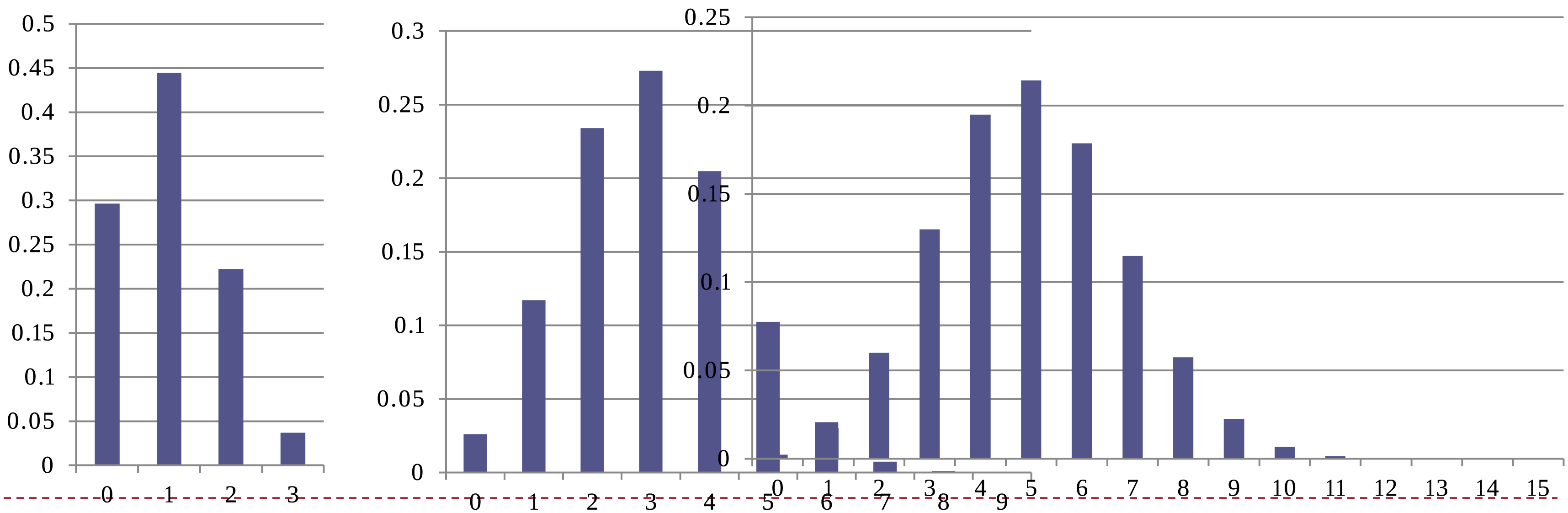
標本平均の分布

母平均 μ、母標準偏差 σ の母集団から大きさ n の無作為標本を抽出するとき、その標本平均 $\bar{x}$ は、n が十分大きければ、近似的に正規分布 $N(\mu, \sigma^2/n)$ に従うとみなすことができる。

中心極限定理

コイン投げの例

- 表が出た時を1、裏が出た時を0として、コインを1～15個投げたときの平均（表が出た回数）をEXCELで計算



推定量

- 母数を推定するために利用される統計量
  - 標本平均:母平均を推定するために利用される推定量
  - 標本分散:母分散を推定するために利用される推定量

推定と検定の基礎

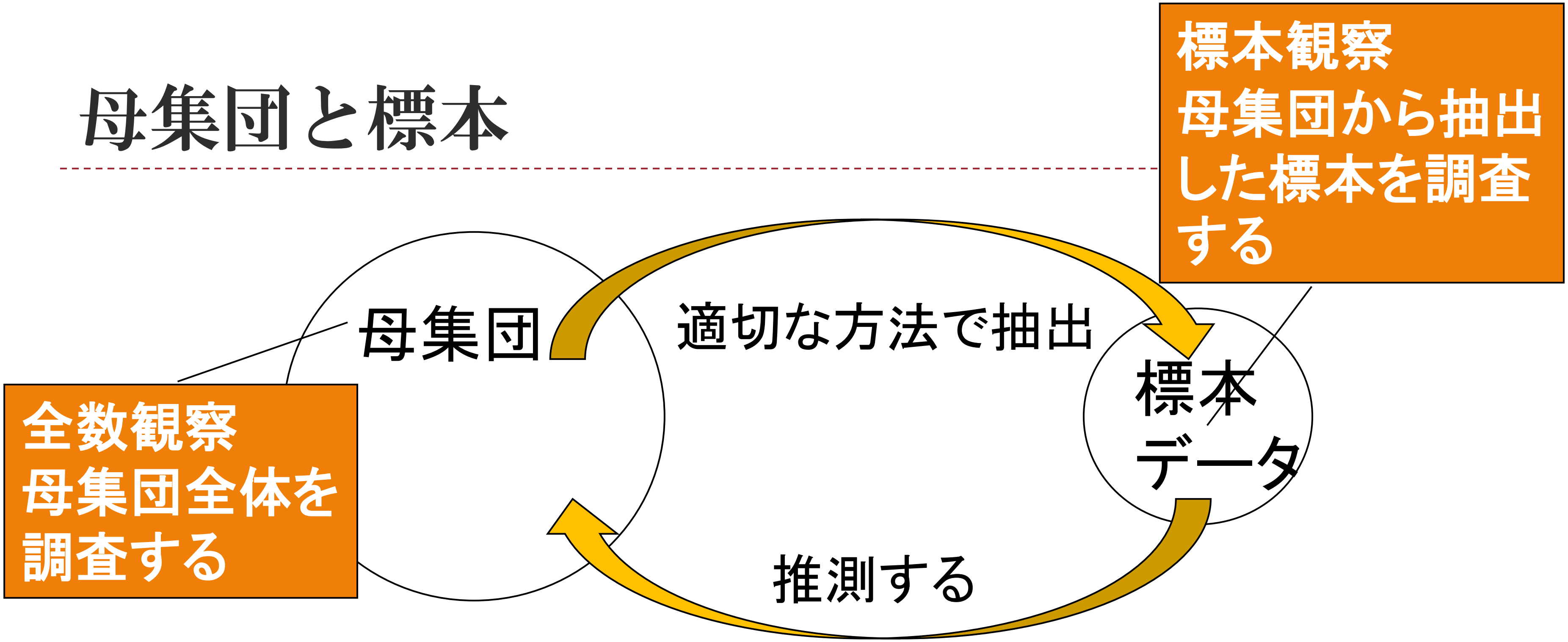
２つの「統計学」

- 記述統計学
  - データを取ってくる
  - データから表やグラフを作成する
  - データから平均や標準偏差を計算する
- 推測統計学
  - データから、母集団の平均や標準偏差を推測したり、仮説を立てて正しいかどうかを検定したりする

なぜ「推測統計学」？

- 世の中にあるほとんどのデータは、母集団から適切な方法で得られた標本を調べて得られたデータだから
  - 標本のデータセットから得られる統計量を、母集団の統計量としている
  - 標本から母集団を推測すると必ず、誤差が生じる
- 誤差を評価する仕組みが必要

母集団と標本



- 大抵の統計データは「標本観察」の結果
  - 母集団がどうなっているかは分からない
- 得られたデータから母集団を推測する

例（母集団と標本）

- 成功率10%の大変難しい手術がある。ある病院ではこれまで、この手術を9回行ったが、すべて失敗した。10人目の患者にこの手術を適用する時、医師は患者に以上の説明をし、「あなたは幸運です。今度は成功する番です」と言った
- 10人でくじを引いた。作ったくじは1本だけ当たりで残り9本ははずれである。9人がくじを引いたが、全員がはずれであった。最後に残った一人は、くじを引く前に「私は当たりだ」と喜んだ
- 何が、違うの？

例　解説

- くじ引きの例：10人で母集団全体
  - 当たる確率である10%は、母集団全体での判断
- 手術の例：10人は標本
  - 成功率10%は、過去にこの手術を実施した患者の集団（数千人）を母集団として計算されたもの
- 標本についての観測値から結論を出すのはそれほど単純ではない

推測統計学で行うこと

- 観測されたデータから、母集団についての情報（母集団の平均、分散、標準偏差の値など）を得、今後に役立てる
  - 推定
  - 検定
  - 予測

統計的推論とは何か

- 人間の推論方法には2つの型がある
  - 帰納法：部分から全体を推定する（昨日までずっと、何千年にもわたって太陽は昇り続けてきたのだから、明日もきっと太陽は昇ってくるだろう）
  - 演繹法：全体から部分を推定する（全ての人間は必ず死ぬ、だから自分も死ぬ）
- 統計的推論（推定、検定）は「帰納的推論」で行われている
  - 自然であるけれども、絶対正しいとはいえない

統計学（統計科学）とは

- いくつかの対象について、観測、調査、実験などを行って得た結果を、数字として表したものを「データ」という
- （不確実性を含む）データから必要な情報を引き出すことを「統計データ解析」といい、そのための方法を「統計データ解析法」という
- 「統計学」とは、統計的方法の体系化ならびに統計的概念の本質に関する研究を課題にしている学問分野である

分布と統計的推論

- もしも、注目している不確実現象に何らかの分布が仮定できるならば、その分布の性質を利用して、何らかの推測（推定、検定）が可能なのではないだろうか？

統計的推論の出発点

正規分布や二項分布など、性質の分かっている分布を「モデル」に使う

不確実性をどうするか？

「100%正しい」
統計的推測は不可能

- 正規分布で考える
- 標準偏差も考慮に入れる
- 幅を持って推定する（区間推定）
- 間違う可能性を考慮に入れる（有意水準）
- 検定の非対称性
  - 帰無仮説が「棄却されない」≠帰無仮説を採択

推定

　母集団から標本をとって得られたデータから、母集団の平均や分散の値（母数）について、それはいくらか、あるいはいくらからいくらぐらいであるかを知ろうとする問題

統計での関心：母集団での平均や分散（真の値）

点推定と区間推定

- 点推定

　平均や分散などの「母数」を、標本観察から得られる一つの値で推定すること

- 区間推定

　点推定より得られた母数の推定が正規分布に従うとして、母数の真の値がある一定の確率（信頼度）で入っていると思われる範囲を、母数の推定値の分散を用いて表す推定。推定された区間のことを「信頼区間」という

推定の例

- 点推定
  - 内閣支持率は34.5%です
- 区間推定
  - 内閣支持率は30.3%から38.8%の間にあると推測されます
  - 但し、その範囲に入らないことも5%の確率であり得ます

検定

　母集団の平均や分散などについて何らかの予想を持っているときに、それを「仮説」という形で提示して、その真偽を標本に基づいて検証すること

検定の例

- さいころを1000回振ったとき、出た目の平均は3.319だった
- ゆがみのない、正確なさいころでの平均は3.5なのだが、このさいころは正確ではないのだろうか？
  - 標本平均は母平均とは完全には一致しない
  - 0.2くらいずれているだけだから、まあ正確？
  - 基準が感覚的だと、判断しづらい
  - 標準偏差も加味して、正確かどうかを考える

片側検定と両側検定

両側検定　帰無仮説：$\mu = \mu_0$、対立仮説：$\mu \neq \mu_0$

片側検定　帰無仮説：$\mu = \mu_0$、対立仮説：$\mu < \mu_0$ 又は $\mu > \mu_0$

- 両側検定：棄却域が両側にある
- 片側検定：棄却域が左側か右側、すなわち片側にしかない検定
- 「対立仮説」の違い

帰無仮説と対立仮説

- 帰無仮説（統計的ゼロ仮説）：検定をする仮説。否定されることを予想している
- 対立仮説：帰無仮説に対立する仮説。帰無仮説の補集合であることが多い

検定には帰無仮説と対立仮説が必要

検定の手順

- 母集団を定義する（母集団分布）
- 帰無仮説と対立仮説を立てる
- どの「検定」を行うか選択する
- 有意水準を決定する
- 標本のデータから検定統計量の値を求める
- 検定統計量の値が棄却域に入っているかどうかを調べる
- 検定統計量の値が棄却域に入っていれば、「対立仮説は正しい」との結論を下す（帰無仮説を棄却）
- そうでなければ、「帰無仮説は誤っているとは言えない」との結論を下す

実際にやってみる

『交際相手の浮気を証明する』

　最近、付き合っている相手が冷たくなったように感じる。

「浮気してない？」って直接尋ねても、本当のことを言うわけない。

統計学的に証明してみる

「交際相手が浮気してるか？」の検定１

- まずは、交際相手を信じる。大前提は『交際相手は浮気をしていない』とする。
- 『交際相手は浮気をしていない』という前提で、その人と周りの現実をよくよく観察してみる。
- すると、デートに誘っても「忙しい」の返事ばかり。デート中もかかってきたケータイにそわそわ。やっぱり、どうもあやしい。

帰無仮説

データの収集

検定統計量の作成

「交際相手が浮気してるか？」の検定２

- 全ての状況証拠を集めて、「浮気をしていない」のにこんな状況になるのは、どのくらいの割合なのだろうと考える
- 全ての状況証拠を集めると、浮気をしていない人がこんな状況になることはとても珍しい。そう判断すれば、「浮気をしていない」という大前提を棄却する
- 「交際相手は浮気をしている」と結論づける

値が棄却域に入っているかどうか

帰無仮説の棄却

仮説検定における注意

　少ないながらも、間違った判断を下してしまう危険性がある。

- 第１種の誤り：帰無仮説が正しいのに、これを否定する過誤（実際には交際相手は浮気していないのに、「浮気している」と結論）
- 第２種の誤り：帰無仮説が間違っているのに、それを否定しない過誤（実際には交際相手は浮気しているのに、「浮気していない」と結論）

どれくらいの確率で「小さい」といえるのか？

第１種の誤り α と第２種の誤り β は互いに関連する。

（一方が小さくなれば他方が大きくなる）

第１種の誤り（有意水準）α を

$\alpha = 0.05$ 又は $\alpha = 0.01$ にし、

その上で β をできるだけ小さくするように

検定を行うのが慣例。

β が十分小さくなるように、標本の大きさ n を決める。

推定や検定するもの

- 母集団のパラメータ（母数）
- 大体において、母集団の平均や分散・標準偏差、比率を推定・検定することが多い
  - 平均
  - 分散・標準偏差
  - 比率
  - 相関、独立性

平均に関する推定

母平均と標本平均

- 母平均：母集団データから計算される平均
- 標本平均：標本データから計算される平均
- 標本平均から母平均を推定・検定する

$$\text{母平均 } \mu : \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

$$\text{標本平均 } \bar{x} : \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

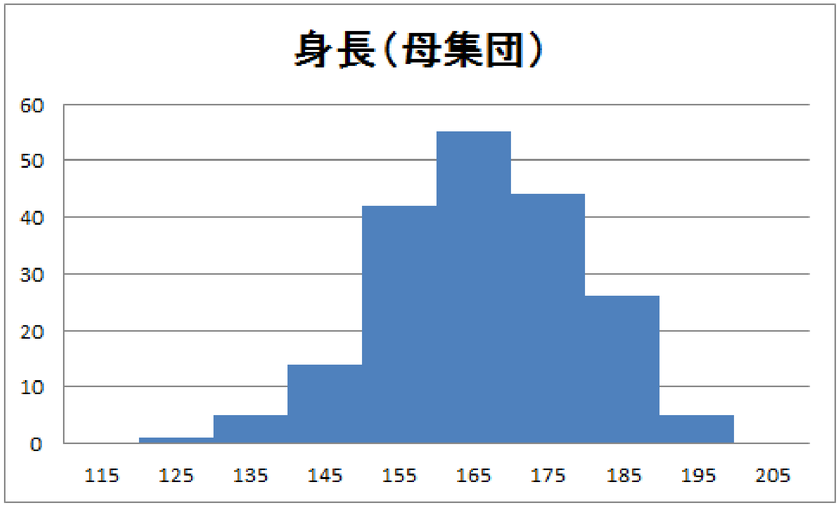
標本平均の性質

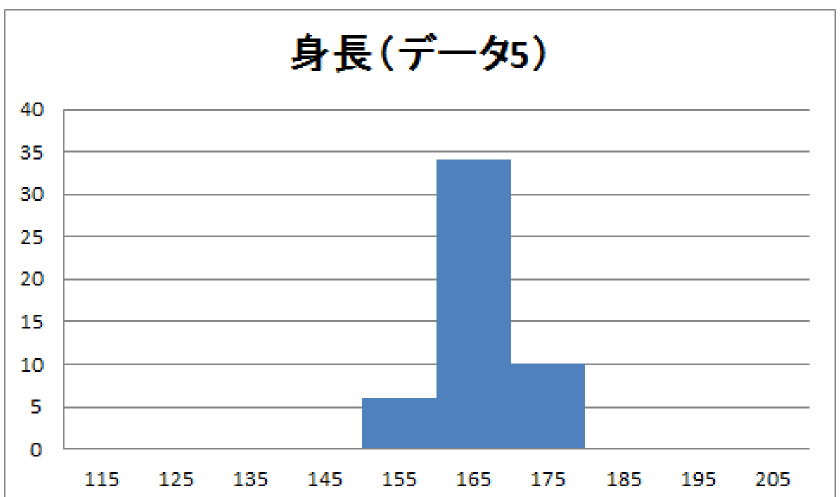
- 見た目（計算公式）は、母平均と同じ
- 標本データから計算される
  - 標本の取り方によって違う結果が出る
  - でも、その期待値は「母平均」になるだろうと予測される
- 同じ母集団から何度も標本をとってきたと仮定して考えると、標本平均は「確率変数である」と考える事ができ、標本平均の「平均」や「分散」を計算する事が可能

個々のデータと平均

- 個々のデータ：ばらつきがある、どのような分布になっているのか分かりにくい（仮定はする）
- 平均を取ると：ばらつきが（個々のデータのばらつきよりも）小さくなる、分布が（大体）分かる
- よって、二つのグループの違いを比較する時には、平均で考える

標本平均の平均と分散の性質例





- 上：192人の身長
- 下：上のデータを母集団として5つのデータを無作為抽出し、平均を計算することを100回繰り返したデータ
- 標本平均の平均は母集団の平均と同じところにある
- 分散は、母集団の分散よりも小さくなっている

母分散と標本分散

- 母分散：母集団データから計算される分散
- 標本分散：標本データから計算される分散
- 標本分散から母分散を推定・検定する
  - 母標準偏差と標本標準偏差でも事情は同じ
    - 分散の正の平方根が標準偏差

$$\text{母分散 } \sigma^2 : \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 = \frac{(x_1-\mu)^2+\cdots+(x_N-\mu)^2}{N}$$

$$\text{標本分散 } s^2 : \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{(x_1-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1}$$

標本分散の性質

- 「母平均」ではなく、「標本平均」からの偏差平方和を計算している
- 「データ数」ではなく、「データ数－１」で割っている
  - 不偏分散
- 標本平均の時と同じく「確率変数」とみなすことができ、標本分散の「平均」や「分散」を計算することが可能

標本平均の分布（標本分布）の性質

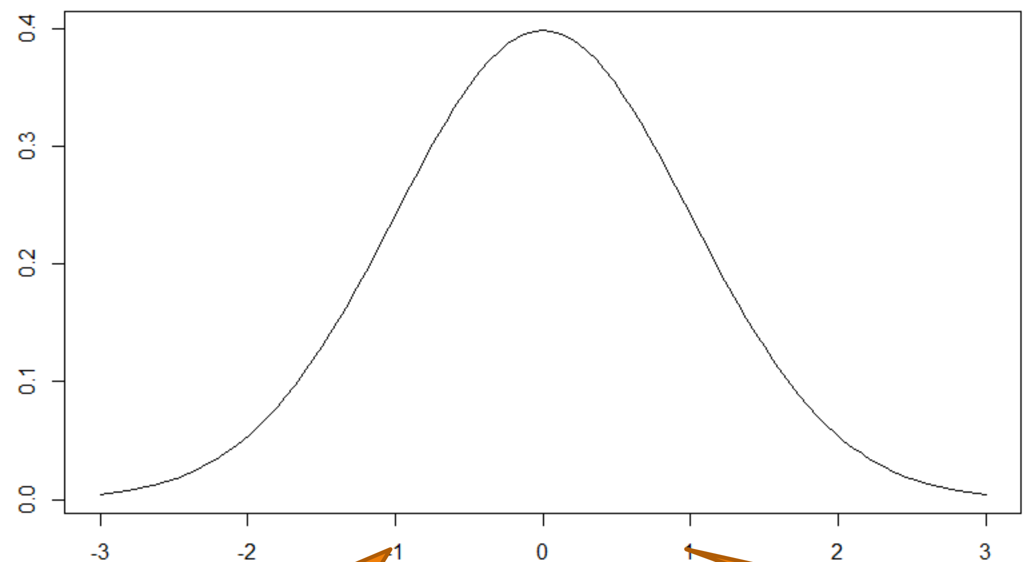
- 平均は、母平均と同じ
- 分散は、母分散をデータ数で割ったもの
- データ数nが大きいとき（大標本）は、正規分布になる
- 母集団が正規分布であれば、データ数nの値に関係なく正規分布に従う

とりあえず、正規分布の性質
を使うと何とかなりそう

正規分布の性質

- 左右対称、山型
- 山の頂上に対するXの値が平均
- 山の中腹に相当するXの値が平均±標準偏差

平均と標準偏差（分散）によって形が決まる



2標準偏差ルール

- 正規分布に従うデータでは、平均から±2標準偏差の範囲内に収まるデータが約95%
- 例：100点満点のテスト
  - 平均は70、標準偏差が5
    - このテストを受けた人の95%は60～80点を取った
    - 80点以上取った人は全体の2.5%しかいない
    - 60点も取れなかった人は全体の2.5%
  - 平均は70、標準偏差が10
    - このテストを受けた人の95%は50～90点を取った
    - 90点以上取った人は全体の2.5%
    - 50点も取れなかった人も2.5%いる

2 標準偏差ルールを厳密にしてみる

- 正規分布の性質
  - 平均μ±2×標準偏差σの面積が約95%
  - すなわち、平均μ±2×標準偏差σの幅の中にデータの約95%が入っている

- 正規分布の性質
  - 平均μ±1.96×標準偏差σの面積が95%
  - すなわち、平均μ±1.96×標準偏差σの幅の中にデータの95%が入っている

実際に平均を推定してみよう

標本平均 $\bar{x}$ の分布は、平均 μ、分散 $\dfrac{\sigma^2}{n}$ の正規分布

確率 0.95 で $\mu - 1.96\dfrac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 1.96\dfrac{\sigma}{\sqrt{n}}$

確率 0.95 で $\bar{x} - 1.96\dfrac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\dfrac{\sigma}{\sqrt{n}}$

$$\boxed{\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}}$$

母平均の信頼度95%信頼区間

（99%の時は、1.96を2.58にする）

「信頼度」の意味

- 信頼度（信頼係数）95%とは？

標本を複数抽出し、
それぞれの標本について、公式にしたがって信頼区間を作成したとき、信頼区間内に母平均が含まれている確率が95%
（信頼区間を20個作成したとき、
そのうち19個の信頼区間には母平均が含まれている）

信頼区間作成シミュレーション

- https://seeing-theory.brown.edu/
- Chapter 4: Frequentist Inference
- Confidence Interval
- 分布（distribution）と標本の大きさ（sample size）と信頼度（confidence level）を選んで「Start Sampling」ボタンを押すと、抽出された標本から作成された信頼区間に母平均が入っているかどうかとその割合を視覚的に表してくれる

「信頼区間」の意味

- 母平均μは、計算された信頼区間の中のどこかの値である
  - 真ん中あたりにあるかもしれない
  - 区間のはしっこにあるかもしれない
- 母平均は定数なので、実際は、標本によって信頼区間が変化する

信頼度はいくらにすればよい？

- 信頼度は、0.95（95%）以外に0.90や0.99もよく使われます
  - 信頼度0.90の時：1.96の代わりに1.64を使う
  - 信頼度0.99の時：1.96の代わりに2.57を使う
- 信頼度をいくつにしたらよいかは、一概には決められない
  - 信頼度を高く（信頼性を高く）すれば、信頼区間が広くなってしまう。
  - 信頼度を低くすれば、信頼区間は狭くなる
  - ま、状況に応じて使いましょう

信頼度を変えたらどうなる？

- 信頼度を99%にする→1.96を2.58に変える→信頼区間の幅が広くなる
- 信頼度を90%にする→1.96を1.64に変える→信頼区間の幅が狭くなる
- 信頼度が上がれば信頼区間の幅が広くなり、信頼度を下げれば信頼区間の幅が狭くなる

母分散が未知の場合は？

- 「母分散が既知の場合」はかなり特殊
- 大抵の場合、母分散は未知
- なので、標本分散で代用
  - n-1で割る方
- 標本のデータの大きさnが十分大きい場合（n=30以上）は普通に正規分布を使った区間推定で推定・検定できる

例

- さいころを1000回ふり、そのとき出た目の平均（標本平均）と標準偏差（標本標準偏差）を求めたところ、平均が3.319、標準偏差が1.713となった
- 出た目の平均の95%信頼区間を求めよ

例解答

区間推定の公式に $\bar{x} = 3.319$、$\sigma = s = 1.723$、$n = 1000$ をあてはめる

$$3.319 - 1.96 \times \frac{1.723}{\sqrt{1000}} \leq \mu \leq 3.319 + 1.96 \times \frac{1.723}{\sqrt{1000}}$$

$$3.319 - 0.107 \leq \mu \leq 3.319 + 0.107$$

$$3.212 \leq \mu \leq 3.426$$

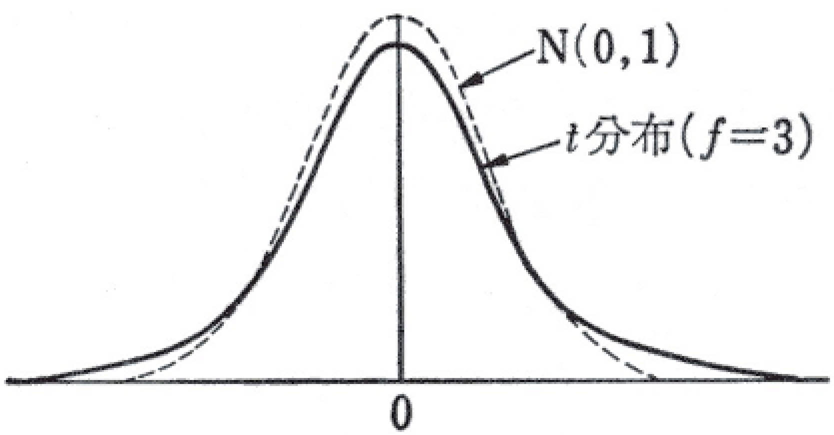
標本の大きさが小さい場合は？

- 標本の大きさが小さい場合には、その標本から計算される分散・標準偏差は、母集団の分散・標準偏差とかけ離れている場合が多い
- 母集団が正規分布に従うと仮定されるとき、母標準偏差が未知の場合でかつ標本の大きさが小さい場合には、正規分布ではなくt分布を使って推定・検定を行う

Studentのt分布

$z = \dfrac{\bar{x} - \mu}{\sqrt{\sigma^2/n}}$ の標準偏差 σ を推定値 s で置き換えた統計量

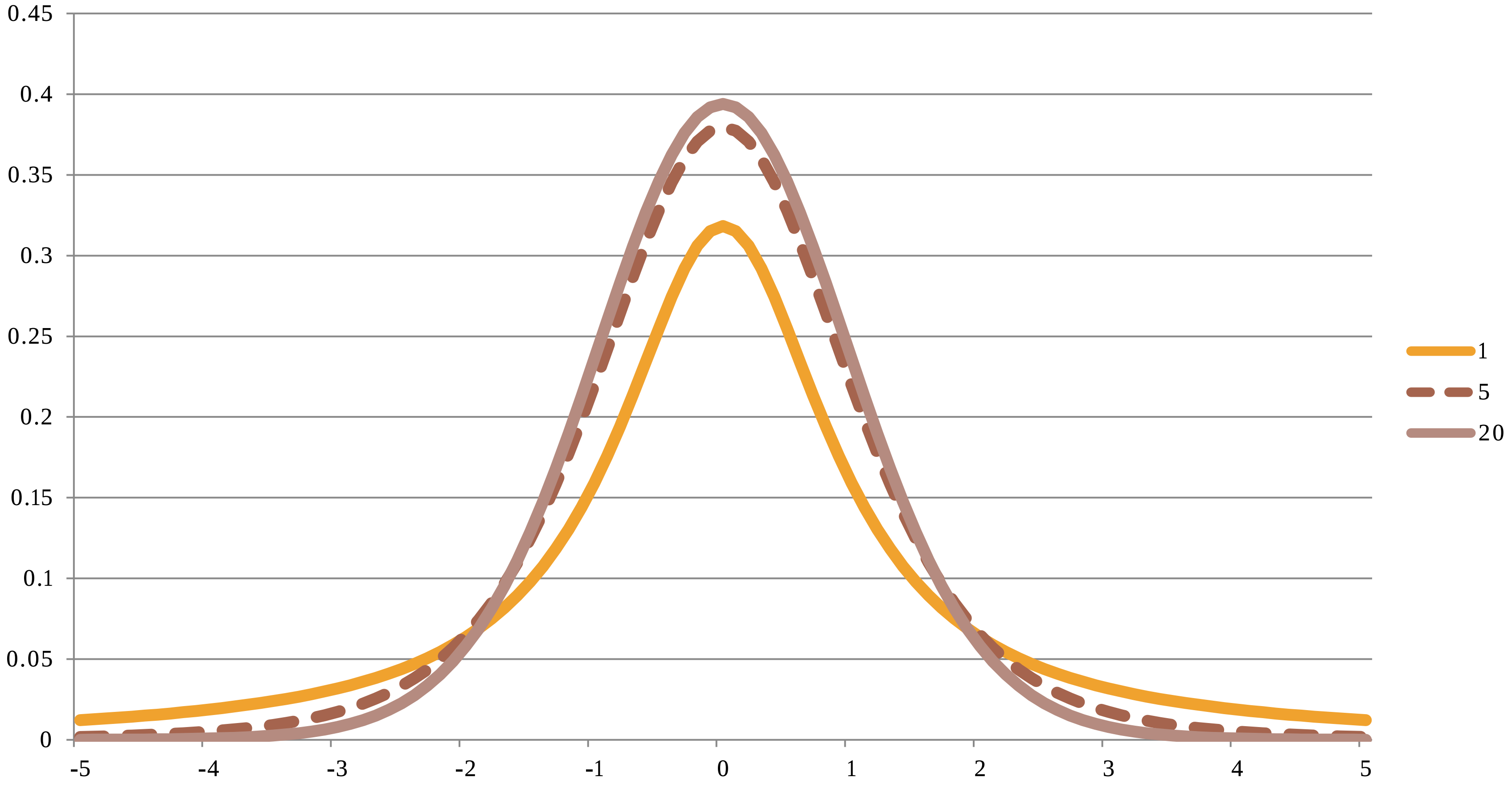
$t = \dfrac{\bar{x} - \mu}{\sqrt{s^2/n}}$ は、$N(0,1)$ の分布より裾を長くひいた分布となる。



Studentのt分布

s^2 の自由度 f（この場合は $f = n - 1$）によって分布が決まる

Studentのt分布のグラフ



t分布の性質

- 正規分布のような、左右対称な山型をしたヒストグラムが描ける
- 正規分布よりはやや緩い山型になる
  - 頂上がやや低い
  - 裾野はやや高い
- 自由度により、分布の形が決まる
  - 推定・検定には「t表」が必要
  - EXCELのT.INV、T.INV.2T関数で計算できる

平均の推定（小標本）

- 小標本＋母分散が未知の場合も基本的に、大標本の時と同じように推定できる
- t分布は自由度によって形が変わるので、1.96のところをt分布における値（自由度n-1のt分布の両側5%点）に置き換える必要がある
  - EXCELで計算可（T.INV.2T(0.05,n-1)）

$$\boxed{\bar{x} - t(p, n-1)\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t(p, n-1)\frac{s}{\sqrt{n}}}$$

自由度n-1のt分布の両側5%点

標本標準偏差

母平均の95%信頼区間

実際に平均を推定してみよう（例）

全長	平均	20.025	=AVERAGE
19.9	分散	0.075833	=VAR.S
19.7			
20.3	自由度	3	=4−1
20.2	t分布の両側5%点	3.182446	=T.INV.2T(0.05,3)
	上側	20.4632	=平均＋t分布の両側5%点×標準偏差／√データ数
	下側	19.5868	=平均－t分布の両側5%点×標準偏差／√データ数

- 母平均の95%信頼区間は、$19.587 \leqq \mu \leqq 20.463$

平均の推定に関するEXCEL関数

- TINV(両側確率,自由度)
- T.INV.2T(両側確率,自由度)
- CONFIDENCE.T(有意水準,標準偏差,標本数)
- CONFIDENCE.NORM(有意水準,標準偏差,標本数)

比率に関する推定

何が必要か

- 平均の推定では「正規分布」を使った
  - 標本平均は、標本の大きさが十分大きければ正規分布に従っている
- 比率の推定に関しても、比率がどの分布に従っているかが分かれば、何とかなりそう
  - 正規分布に従っていればなお良し

母比率と標本比率

- 比率：全体のうち、ある特性を持つ個体の集まり
  - 高血圧の人の割合
  - 内閣支持率
  - 製品の不良率
- 母比率R：母集団における比率
- 標本比率r：標本データから計算される比率

特性を持つを「1」、持たないを「0」で表すと、標本比率rは標本平均とみなすことができる。

比率の分布（標本分布）の性質

比率（割合）r＝起こった回数X／全体n

二項分布

- 平均は、母集団における比率Rと同じ
- 分散は、母集団分布の分散nR(1-R)をデータの大きさnの2乗で割ったR(1-R)/n で計算できる
- データの大きさnが大きいとき（大標本）は、中心極限定理により正規分布になる
  - nが小さい時は二項分布

母比率の推定

nが十分大きければ、標本比率rは母比率R、分散$\dfrac{R(1-R)}{n}$、標準偏差$\sqrt{\dfrac{R(1-R)}{n}}$の正規分布に従う。

$$r - 1.96\sqrt{\frac{R(1-R)}{n}} \leq R \leq r + 1.96\sqrt{\frac{R(1-R)}{n}}$$

（信頼度95%）

注：左辺と右辺の母比率Rは標本比率rで置きかえる。

例題 1

　ある工場の製品400個について検査したところ、不良品が18個あった。全製品における不良率を、信頼度95%で推定せよ。

母比率の推定（例）

不良率：$p = 18/400 = 0.045$

$$0.045 - 1.96\sqrt{\frac{0.045(1-0.045)}{400}} = 0.025$$

$$0.045 + 1.96\sqrt{\frac{0.045(1-0.045)}{400}} = 0.065$$

なので、信頼度 95 %の信頼区間は

$$0.025 \leq P \leq 0.065$$

標本の大きさ１

標本の大きさが50で、商品名を知っていると答えた人が20人いた場合

点推定値　$P = 20/50 = 0.40$

区間推定　$P \sim 0.40 \pm 1.96\sqrt{0.40(1-0.40)/50}$

$0.264 < P < 0.536$

標本の大きさ２

標本の大きさが500で、商品名を知っていると答えた人が200人いた場合

点推定値 $P = 200/500 = 0.40$

区間推定 $P \sim 0.40 \pm 1.96\sqrt{0.40(1-0.40)/500}$

$0.357 < P < 0.443$

割合の推定においては、標本の大きさが小さいと信頼区間の幅が広くなってしまう。

標本はたくさんあった方が良いのか？

$$r - 1.96\sqrt{\frac{R(1-R)}{n}} \leq R \leq r + 1.96\sqrt{\frac{R(1-R)}{n}}$$

r	0.5	n	下限	上限	差	
		10	0.1901	0.8099	0.6198	
		20	0.2809	0.7191	0.4383	0.181537
		100	0.4020	0.5980	0.1960	
		110	0.4066	0.5934	0.1869	0.009121
		1000	0.4690	0.5310	0.0620	
		1010	0.4692	0.5308	0.0617	0.000308
		10000	0.4902	0.5098	0.0196	
		10010	0.4902	0.5098	0.0196	9.79E-06

- ある程度（10000くらい）大きくなると、上限と下限の差の減少カーブが緩くなる（差があまり変わらなくなる）

小標本の場合

- 二項分布で考える
  - Excel
  - EZRなどの統計解析ソフトウエアを使う

小標本の推定例

- 347例を調査して、11例が当てはまった
  - 点推定:0.0317
  - 95%信頼区間:0.01593～0.05601

π	n	x	下側	上側
0.0317	347	11	0.5793	0.5420
0.013266	347	11	0.9973	0.0074
0.050135	347	11	0.0666	0.9629
0.01593	347	11	0.9892	**0.0250**
0.05601	347	11	**0.0250**	0.9874
下側：	確率πの2項分布でx以下になる確率			
上側：	確率πの2項分布でx以上になる確率			
上側、あるいは下側確率が0.025になるように				
πの項を調整（ゴールシーク利用）				
（データ→What-If分析→ゴールシーク）				

平均に関する検定

統計的仮説検定とは？

- 母集団の平均や分散などについて何らかの予想を持っているときに、それを「仮説」という形で提示して、その真偽を標本に基づいて検証すること

例

- ある水族館にいるたこさんは、サッカーの試合でどちらのチームが勝つか当てることができると言われている
- 以下の場合、このたこさんはサッカーの勝利チームを正しく当てることができていると言えるだろうか？
  - 8試合中8試合全部当てた
  - 8試合中7試合当てた
  - 8試合中6試合当てた

例

シミュレーション＋2項分布を利用
（rand関数、binom.dist関数）

- 「たこさんには勝利チームを当てる能力はない（勝つチームをランダムに選んで、選んだそのチームがたまたま勝っただけ）」という帰無仮説のもとで、それぞれが起こる確率を計算してみる
  - 8試合全部当てた
  - 8試合中7試合当てた
    - 7試合当てるのが「まれ」→8試合当てるのも「まれ」
  - 8試合中6試合当てた
- 確率が「まれ」だったら帰無仮説を棄却

検定に関する用語

- 帰無仮説：検定する仮説
- 対立仮説：帰無仮説に対立する仮説、帰無仮説の補集合
- 有意水準（危険率）：帰無仮説を棄却する際の『極めて稀』の基準確率、帰無仮説が正しいのに間違って帰無仮説を棄却してしまう確率

平均に関する検定

- 推定の時と同じ仮定を使う
  - 大標本
  - 大数の法則と中心極限定理により、正規分布で近似できる
- 『帰無仮説』が必要
- 有意水準の決定（5%にするか、それとも……）
- 検定統計量の計算
- p値（有意確率）
- 棄却域：検定統計量を計算した結果、その値がその領域に入った場合、帰無仮説を棄却する領域

標本平均の分布（標本分布）の性質

- 平均は、母平均と同じ
- 分散は、母分散をデータ数nで割ったもの
- データ数nが大きいとき（大標本）は、正規分布になる
- 母集団が正規分布であれば、データ数nの値に関係なく正規分布に従う

平均の推定

標本平均 $\bar{x}$ の分布は、平均 μ、分散 $\dfrac{\sigma^2}{n}$ の正規分布

確率0.95で $\mu - 1.96\dfrac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 1.96\dfrac{\sigma}{\sqrt{n}}$

確率0.95で $\bar{x} - 1.96\dfrac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\dfrac{\sigma}{\sqrt{n}}$

$$\boxed{\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}}$$

母平均の信頼度95%信頼区間

（99%の時は、1.96を2.58にする）

検定の方法

- 帰無仮説(母平均についての予想)と対立仮説を決める
- 帰無仮説が正しいという判断のもとで、母集団から得た標本・統計量の分布を定める(大抵は正規分布(平均と分散が必要))
- 有意水準(危険率)を決める(たいていの場合5%)
- 対立仮説に有利となる棄却域を設定する
- 標本を抽出する
- 標本から計算される統計量(実測値)が棄却域にあれば、帰無仮説になることは「まれ」だということが分かるので、帰無仮説を棄却し、対立仮説を採択する
- 統計量が棄却域に無ければ、帰無仮説はあり得るということで判断保留

平均の検定

標本の大きさを n、母標準偏差 σ は既知とする。

もし帰無仮説 $\mu = \mu_0$（μ_0 は数値）が正しければ、標本平均 $\bar{x}$ は平均 μ_0、標準偏差 $\dfrac{\sigma}{\sqrt{n}}$ の正規分布に従う。

確率 0.95 で $\mu_0 - 1.96\dfrac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu_0 + 1.96\dfrac{\sigma}{\sqrt{n}}$

標本平均がこの外に有れば「まれ」

$$\boxed{\bar{x} < \mu_0 - 1.96\frac{\sigma}{\sqrt{n}},\ \ \mu_0 + 1.96\frac{\sigma}{\sqrt{n}} < \bar{x}}$$

有意水準5%の棄却域（両側検定）

有意水準はいくらにすればいいの？

- 有意水準には0.05（5%）がよく使われる
  - 誤る可能性を少なくしたい場合は有意水準を小さく、誤りをより多く認めても構わない場合には有意水準を大きくする
  - 有意水準を小さくすると、誤る可能性は低くなるが仮説が棄却されにくくなるので、結果として結論が出にくくなる
  - 但し、棄却されないからといって勝手に有意水準を変えてはいけない

母分散が未知の場合は？

- 「母分散が既知の場合」はかなり特殊
- 大抵の場合、母分散は未知
- なので、標本分散で代用
- 標本のデータの大きさnが十分大きい場合（n=30以上）は普通に正規分布を使った区間推定で検定できる
- 標本データの大きさnが大きくない場合は、別の考察が必要

例題（検定）

- さいころを1000回ふり、そのとき出た目の平均（標本平均）と標準偏差（標本標準偏差）を求めたところ、平均が3.319、標準偏差が1.713となった
- 出た目の平均が3.5であるかどうかを検定せよ

解答１（検定）

帰無仮説　$H_0 : \mu_0 = 3.5$、有意水準　５％（$\alpha = 0.05$）

$n = 1000$、$s = 1.713$ として、

有意水準５％の棄却域は

$$\bar{x} < 3.5 - 1.96 \times \frac{1.713}{\sqrt{1000}}, \quad 3.5 + 1.96 \times \frac{1.713}{\sqrt{1000}} < \bar{x}$$

すなわち、$\bar{x} < 3.39, \quad 3.61 < \bar{x}$

標本平均の値 $\bar{x} = 3.319$ はこの範囲に入っているので、有意水準５％で帰無仮説を棄却、すなわち、このさいころの母平均は 3.5 であるとはいえない、と結論される。

検定の方法２

- 帰無仮説（母平均についての予想）と対立仮説を決める
- 帰無仮説のもとで、対象となる統計量の分布を定める（大抵は正規分布（平均と分散が必要））
- 有意水準（危険率）を決める（たいていの場合5%）
- 標本を抽出して統計量（平均や分散）を計算
- 帰無仮説が正しいとして標準化（検定統計量）
- 検定統計量について、有意水準から分かる分布のパーセント点と比較、あるいは、p値を計算する
- 検定統計量が有意水準から分かる%点より大きい、あるいはp値が有意水準より低ければ、帰無仮説になることは「まれ」だということが分かるので、帰無仮説を棄却し、対立仮説を採択する
- そうでなければ、帰無仮説はあり得るということで判断保留

平均の検定（ある値との比較）

帰無仮説 $H_0 : \mu = \mu_0$

対立仮説 $H_1 : \mu \neq \mu_0$

（μ_0：検定したい値（数字））

この値の絶対値と、
1.96（有意水準5%の場合）
を比較する

検定統計量：$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

$\bar{x}$：標本平均、s：標本標準偏差（$n-1$ で割る方）、

n：データ数

解答2

帰無仮説　$H_0 : \mu_0 = 3.5$、有意水準　5％（$\alpha = 0.05$）

⇒ 検定統計量：

$$t = \frac{3.319 - 3.5}{1.713/\sqrt{1000}} = -3.341$$

検定統計量の絶対値3.341と、
有意水準5%の基準点1.96とを比較して、
検定統計量の絶対値の値の方が大きいので、
帰無仮説を棄却する。

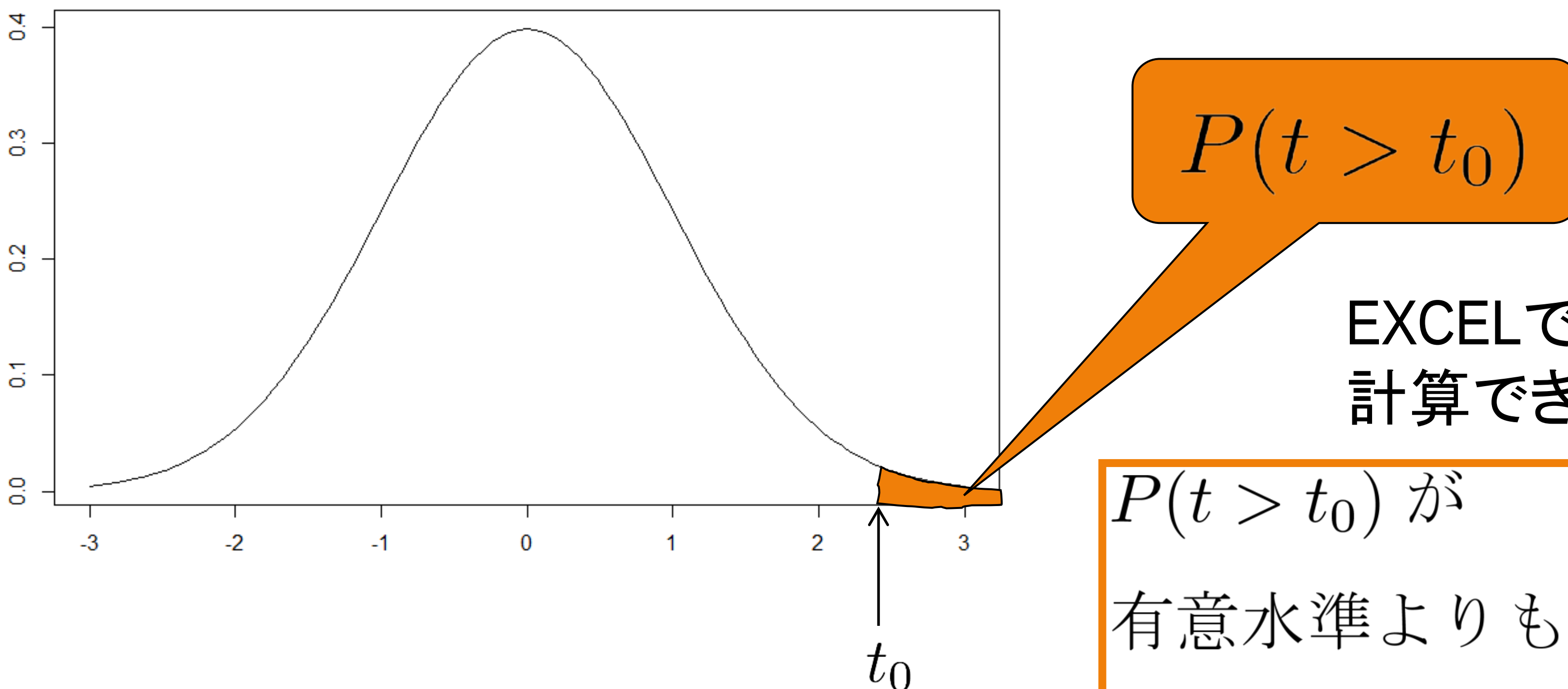
p値（有意確率）

計算機が発達していない頃、検定で棄却するかどうか（有意かどうか）の判断を行う場合には、検定統計量が従っている分布の５％点（あるいは１％点）から棄却域を作成し、標本から求められた検定統計量の値 t_0 が棄却域に入っているかどうかを確かめ、有意かどうかの判断をしてきた。

だが、現在、計算機が発達した為、ある分布において t_0 以上の値をとる確率 $P(t > t_0)$ が簡単に計算できるようになった。そこで、t_0 以上の値をとる確率 $P(t > t_0)$ が５％、あるいは１％より小さい時に「有意である」と判断するという形をとることがある。

この確率 $P(t > t_0)$ のことを p 値（有意確率）という。

p値（有意確率）を図にすると



EXCELで
計算できる

$P(t > t_0)$ が
有意水準よりも
小さければ
帰無仮説を棄却。

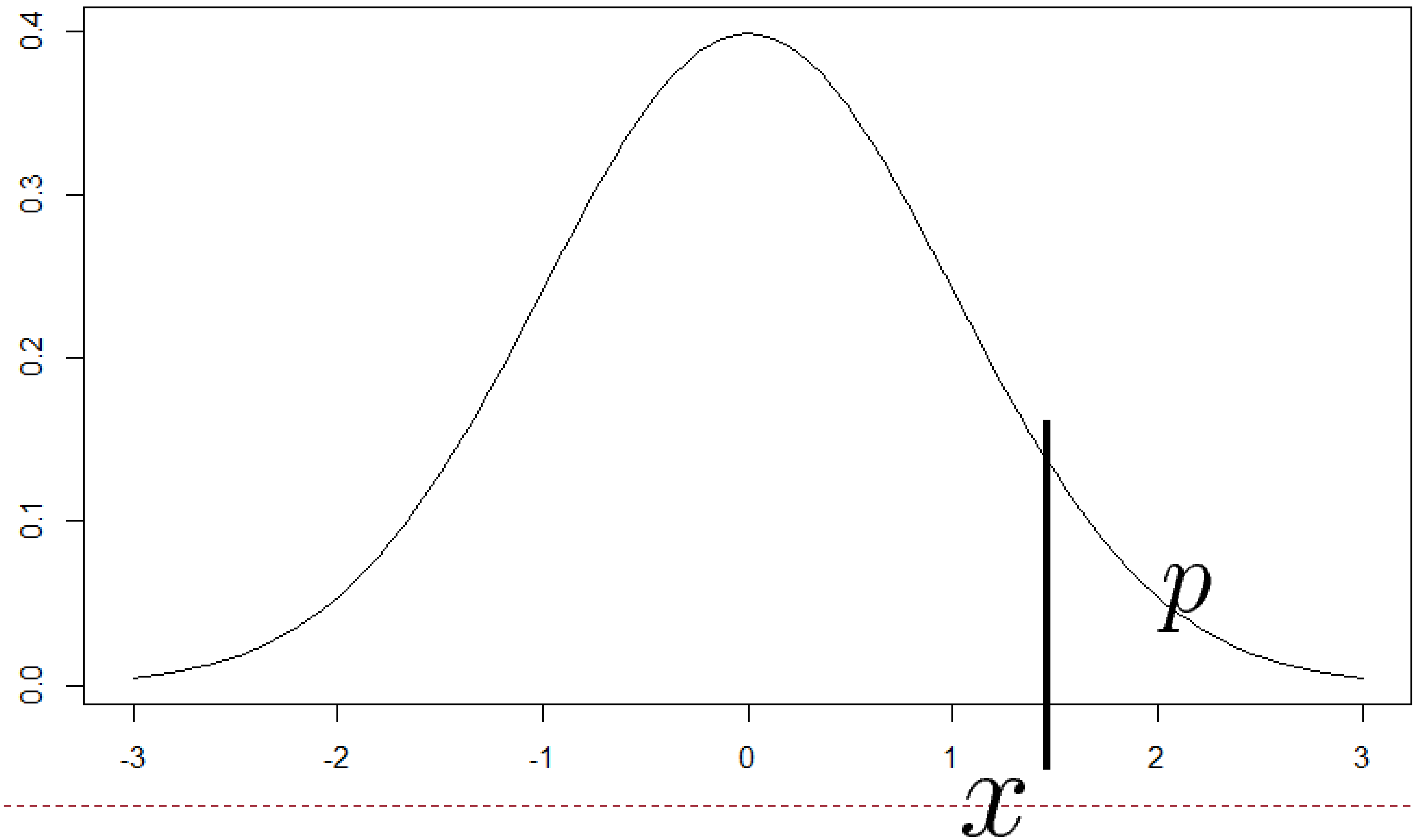
上側、下側、両側p値

- 上側p値：確率分布で、点xの右側を占める確率p
- 下側p値：確率分布で、点xの左側を占める確率p
- 両側p値：確率分布で、点xの右側を占める確率p/2

点xが与えられている時に
確率pを求める

実際は、EXCELで計算（次回以降）

図で表すと



p値の計算に用いるEXCEL関数

- 正規分布：
  - NORM.DIST(統計量,平均,標準偏差,関数形式)
  - NORM.S.DIST(統計量,平均,標準偏差,関数形式)
- t分布：
  - T.DIST(統計量,自由度,両側か片側か)
  - T.DIST.2T(統計量,自由度)
- カイ2乗分布：
  - CHISQ.DIST(統計量,自由度)
  - CHISQ.DIST.RT(統計量,自由度)
- F分布：
  - F.DIST(統計量,自由度1,自由度2)
  - F.DIST.RT(統計量,自由度1,自由度2)
- 二項分布：
  - BINOM.DIST(成功数,試行回数,成功率,関数形式)

p値を使った例題解答

- EXCELの関数NORMDISTを用いて、母平均3.5、標準偏差1.713/√1000におけるx=3.319の累積確率（両側）を求めると、0.000833738となる
- これは、有意水準5%（0.05）よりも遥かに小さいので、母平均が3.5であるという仮説は棄却される、すなわち、母平均は3.5ではない

標本の大きさ	標本平均	標本標準偏差	母平均
1000	3.319	1.713	3.5
0.000833738			

パーセント点

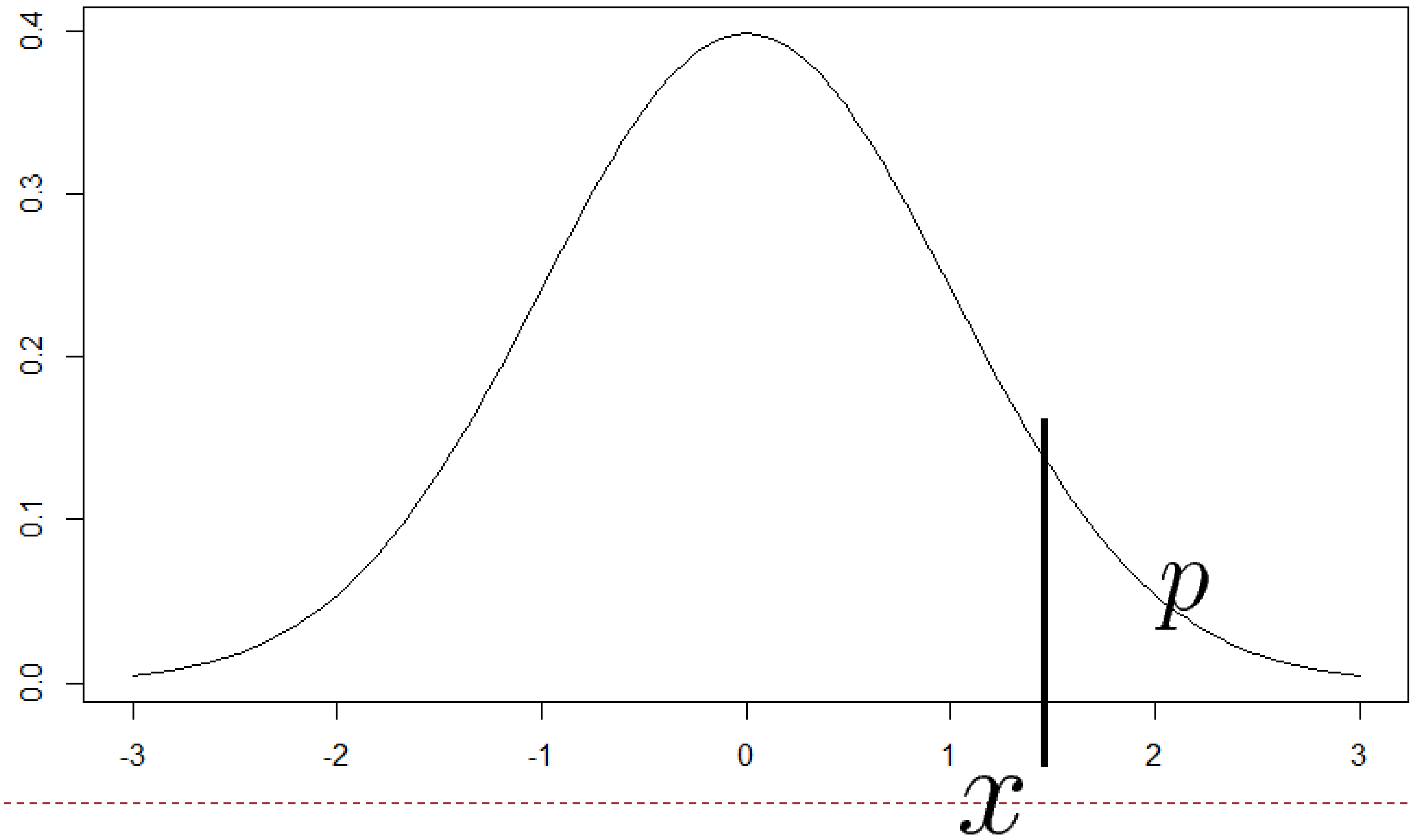
- データを小さい順に並べた時、その値よりも小さな値の割合が指定された割合になる値
  - 10%点：その値より小さいデータが全体の10%になる値
  - 50%点＝中央値
  - 0%点＝最小値、100%点=最大値
  - 上側4分位値＝75%点、下側4分位値=25%点

上側、下側、両側パーセント点

- 上側100pパーセント点：確率分布で、ある点xより右側の確率がpである、点xのこと
- 下側100pパーセント点：確率分布で、ある点xより左側の確率がpである、点xのこと
- 両側100pパーセント点：確率分布で、ある点xより右側の確率がp/2である、点xのこと

確率pが与えられている時に
点xを求める

図で表すと



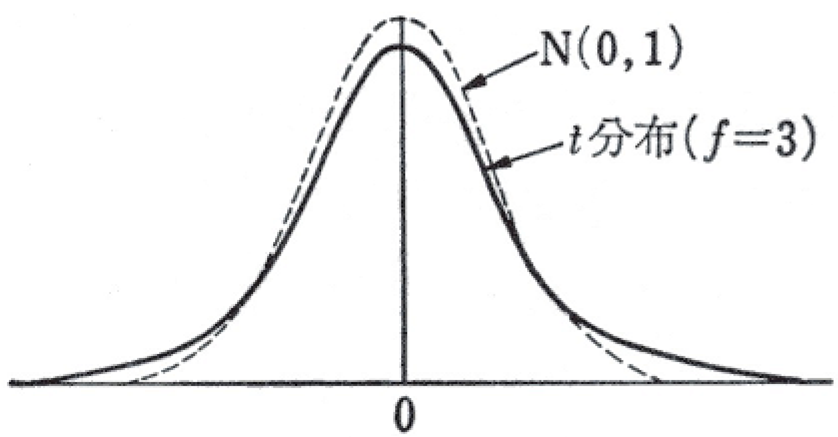
標本の大きさが小さい場合は？

- 標本の大きさが小さい場合には、その標本から計算される分散・標準偏差は、母集団の分散・標準偏差とかけ離れている場合が多い
- 母集団が正規分布に従うと仮定されるとき、母標準偏差が未知の場合でかつ標本の大きさが小さい場合には、正規分布ではなくt分布を使って検定を行う

Studentのt分布

$z = \dfrac{\bar{x} - \mu}{\sqrt{\sigma^2/n}}$ の標準偏差 σ を推定値 s で置き換えた統計量

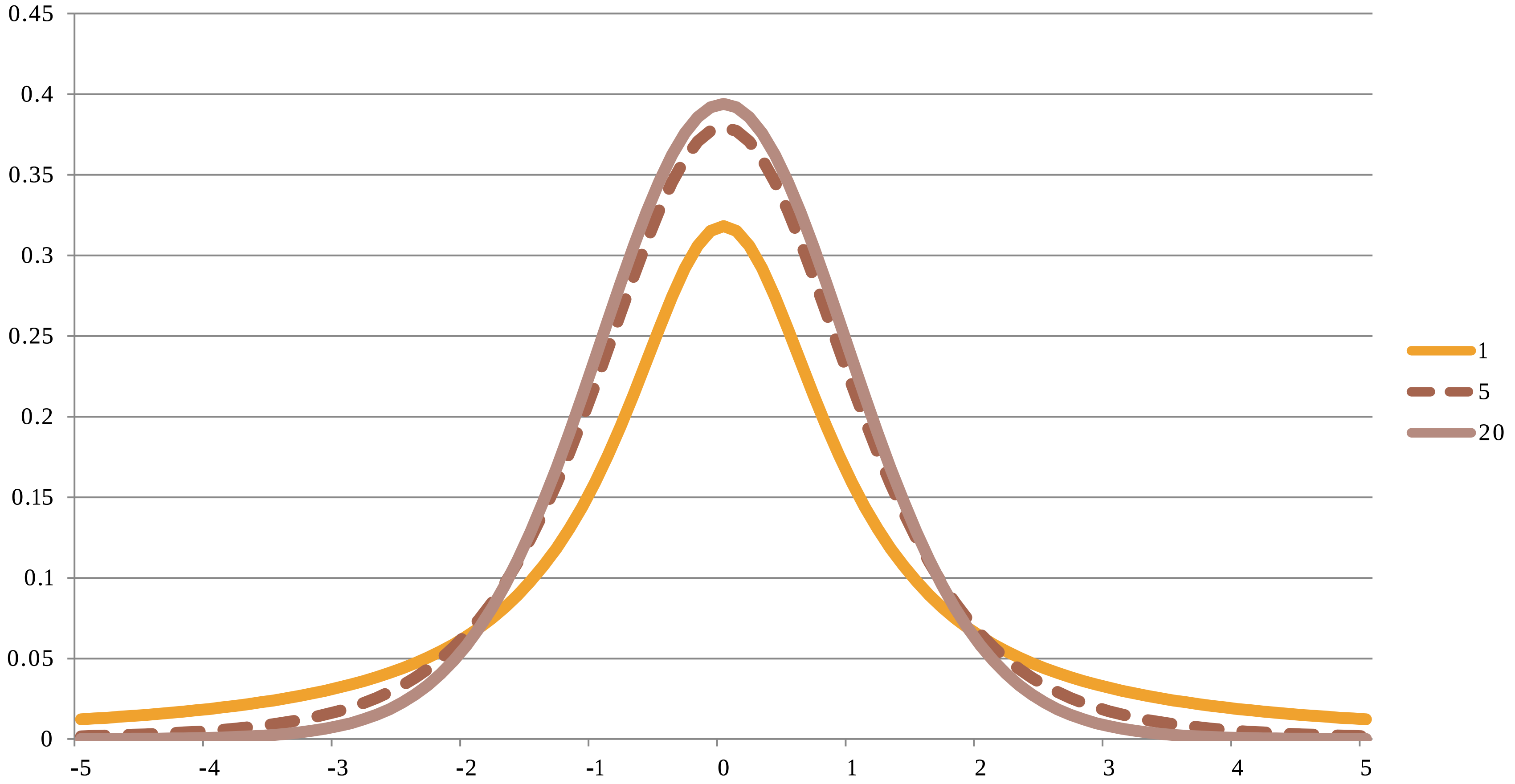
$t = \dfrac{\bar{x} - \mu}{\sqrt{s^2/n}}$ は、$N(0,1)$ の分布より裾を長くひいた分布となる。



Studentのt分布

s^2 の自由度 f（この場合は $f = n - 1$）によって分布が決まる

Studentのt分布のグラフ



t分布の性質

- 正規分布のような、左右対称な山型をしたヒストグラムが描ける
- 正規分布よりはやや緩い山型になる
  - 頂上がやや低い
  - 裾野はやや高い
- 自由度により、分布の形が決まる
  - 推定・検定には「t表」が必要

平均の検定（小標本）

小標本の場合：t分布で検定を行う（t検定）

$$\bar{x} < \mu_0 - t(p, n-1)\frac{s}{\sqrt{n}}, \ \mu_0 + t(p, n-1)\frac{s}{\sqrt{n}} < \bar{x}$$

有意水準5%の棄却域（両側検定）

帰無仮説に従うと仮定して標準化

検定統計量

$T = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$ の絶対値と、

自由度 $n-1$ の t 分布の両側 5 % 点 $t(p, n-1)$ の値とを比較し、T の方が大きければ仮説を棄却。

平均の検定（小標本例）

全長	平均	20.025	=AVERAGE
19.9	分散	0.075833	=VAR.S
19.7			
20.3	帰無仮説	19.89	
20.2	検定統計量	0.980469	=(標本平均−帰無仮説)/標準偏差/√￣標本の大きさ
	自由度	3	=標本の大きさ−1
	t分布の値	3.182446	=T.INV.2T(0.05,3)

帰無仮説:母平均＝19.89
検定統計量の値が自由度3のt分布の両側5%点より小さいので、帰無仮説は棄却できない。
母平均は19.89でないとはいえない。

本当に平均に「違い」があるのか検定してみよう

- 前提：小標本＋セ

帰無仮説 $H_0 : \mu_A = \mu_B$ ⇒ $\mu_A - \mu_B = 0$

対立仮説 $H_1 : \mu_A \neq \mu_B$

この値の絶対値と、
自由度nA+nB-2の場合のt分布の
両側95%点を比較する

帰無仮説に従うと仮定して標準化

⇒ 検定統計量：$t = \dfrac{\bar{x}_A - \bar{x}_B}{s^* \sqrt{1/n_A + 1/n_B}}$

$\bar{x}_A$：群 A の標本平均、s_A：群 A の標本標準偏差、n_A：群 A のデータ数

$\bar{x}_B$：群 B の標本平均、s_B：群 B の標本標準偏差、n_B：群 B のデータ数

$$s^* = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}}$$

共通の標準偏差

2 群の比較（例）

- A市とB市の住民の睡眠時間についての調査
  - 2つの市から無作為に15人ずつ選び出して調査
- 結果
  - A市 : 平均8.27、標準偏差1.16
  - B市 : 平均6.60、標準偏差1.30
- 検定したいこと
  - 睡眠時間について、2つの市の間に差があるかどうか？

帰無仮説 $H_0 : \mu_A = \mu_B$

対立仮説 $H_1 : \mu_B \neq \mu_B$

2群の比較（例解答）

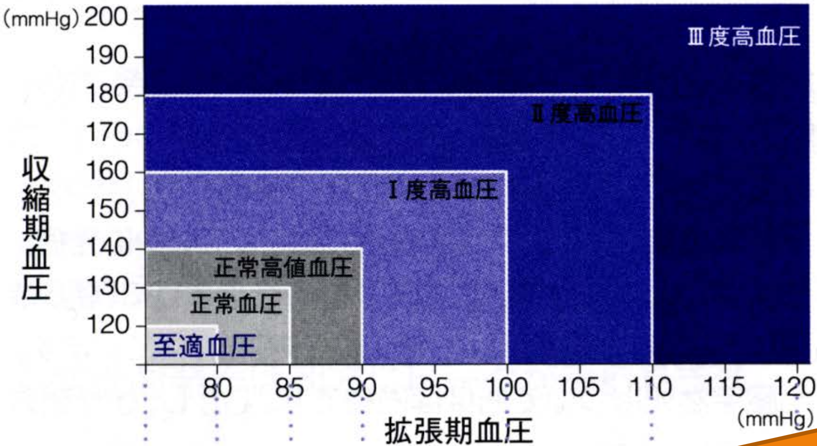
$$s^* = \sqrt{\frac{(15-1)1.16^2 + (15-1)1.30^2}{15+15-2}} = 1.232496$$

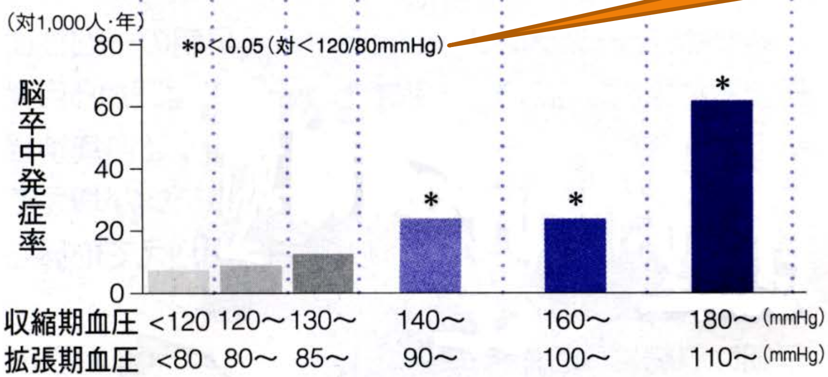
$$\text{検定統計量}: t = \frac{8.27 - 6.60}{1.23\sqrt{1/15 + 1/15}} = 3.703$$

検定統計量の絶対値3.703と、
t分布における自由度28の限界値2.368
（両側5%点）とを比較して、
検定統計量の絶対値の値の方が大きいので、
帰無仮説を棄却する。

p値：0.000925

高血圧の診断と分類





血圧と脳卒中発症率の関係【久山町研究】

【対象】1961年から久山町の循環器健診を受診した住民から設定した第一集団、60歳以上の男女580名。
【方法】追跡32年、性・年齢調整

〔日本高血圧学会. 高血圧治療ガイドライン2009より改変〕

血圧が高くなるにつれて**脳卒中発症リスクは高まります。**
至適血圧を目指しましょう。

この意味は？

EXCELを用いて平均の比較

- T.TEST(A群のデータ,B群のデータ,検定の指定,検定の種類)
- 検定の指定：片側検定なら1、両側検定なら2
- 検定の種類：対なら1、等分散の２標本なら2、不等分散の２標本なら3

分析ツールを用いてもできる（t検定）

片側検定と両側検定

- 両側検定:棄却域が両側にある
- 片側検定:棄却域が左側か右側、すなわち片側にしかない検定
- 「対立仮説」の違い

両側検定　帰無仮説：$\mu = \mu_0$、対立仮説：$\mu \neq \mu_0$

片側検定　帰無仮説：$\mu = \mu_0$、対立仮説：$\mu < \mu_0$ 又は $\mu > \mu_0$

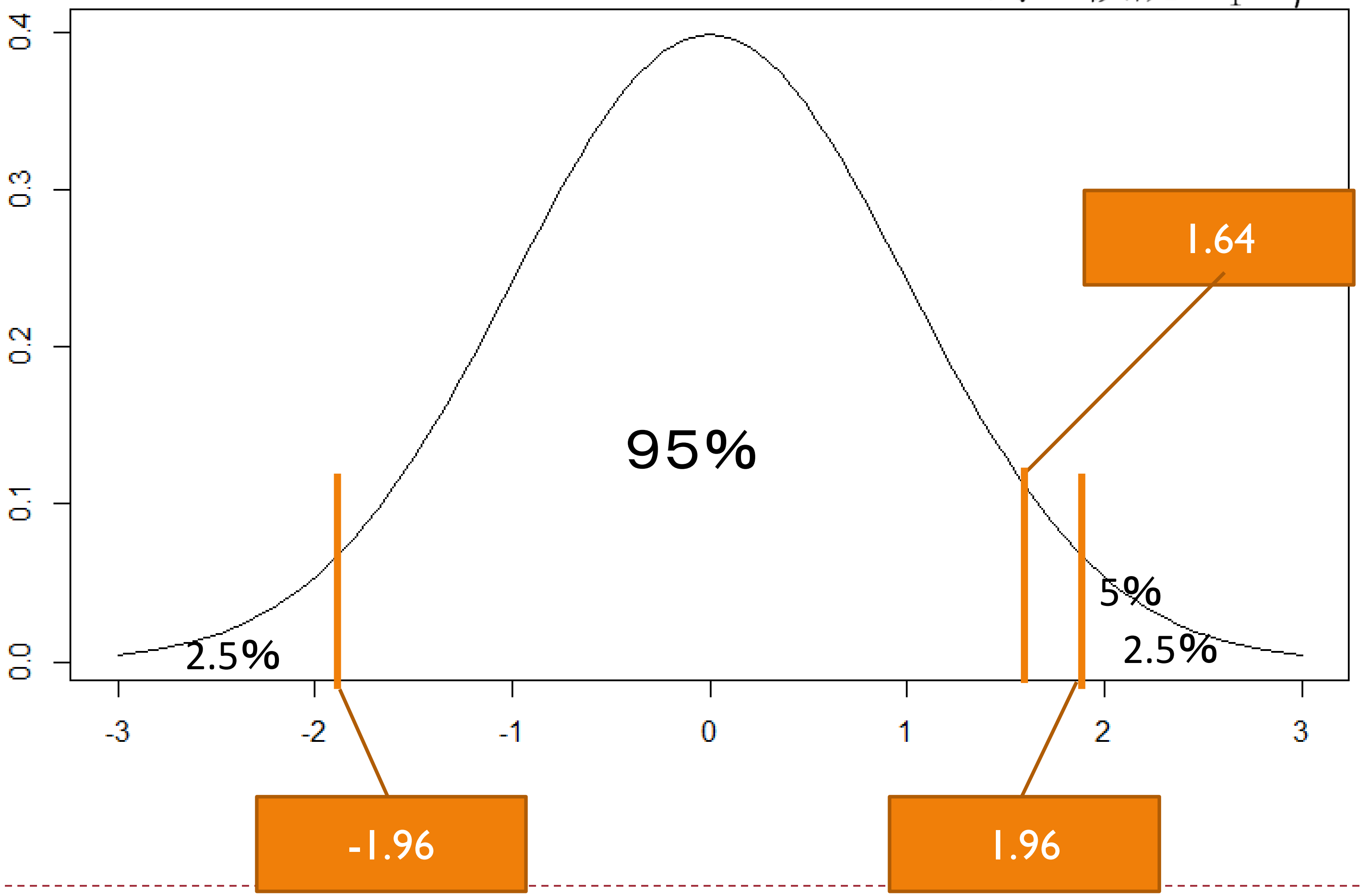
両側検定と片側検定

帰無仮説 $H_0 : \mu = \mu_0$

対立仮説 $H_1 : \mu \neq \mu_0$

帰無仮説 $H_0 : \mu = \mu_0$

対立仮説 $H_1 : \mu > \mu_0$



片側検定の棄却域

- 棄却域が片側にしかないので、その分色々と変わる

対立仮説が $\mu < \mu_0$（左側）の場合：$\bar{x} < \mu - 1.64\dfrac{\sigma}{\sqrt{n}}$

対立仮説が $\mu > \mu_0$（右側）の場合：$\mu + 1.64\dfrac{\sigma}{\sqrt{n}} < \bar{x}$

有意水準5%の場合。
有意水準が1%、10%の場合はそれぞれ、1.64を2.33、1.28に変える

例題

749	934	680	562	814
647	515	839	388	619
674	494	698	763	597
559	668	770	620	551

- 昨年度の勤労者世帯（２人以上）の平均年収は620万円でした
- 今年度の平均年収が620万円より多いかどうかを検定したい
- ただし、母分散は未知であるとする

例題解答

- 母分散が未知＋小標本なので、t分布を用いて検定を行う

帰無仮説：$\mu = 620$
対立仮説：$\mu > 620$

右片側検定

帰無仮説の下では、統計量 $T = \dfrac{\bar{X} - 620}{s/\sqrt{20}}$ は、自由度 19 の t 分布に従う。

有意水準 5 %、右片側検定での t 分布における棄却域は $T > 1.73$。

$\bar{X} = 657.1$、$s = 130.8$ より、$T = \dfrac{657.1 - 620}{130.8/\sqrt{20}} = 1.27$。

この値は棄却域内に入っていないので、帰無仮説は棄却されない。すなわち、今年度の平均年収は 620 万円より大きいとはいえない（保留）。

仮説検定の弱点

- 仮説検定では常に、帰無仮説を設定し、帰無仮説を棄却することで対立仮説が正しいことを示している（後件否定のロジック）

- 後件否定のロジックは，真偽の2値を問う形式論理学では常に論理的に妥当な推論となっているが，仮説検定のような確率を伴う議論においては厳密に適用できないことが分かっている
- 仮説検定では，帰無仮説それ自体の真偽判定は不可能
  - 帰無仮説が棄却できない＝判断を保留
- 対立仮説の方も、直接真偽を確かめているとは言い難い

最近のデータ分析結果に併記されているもの

- p値
- 効果量
- 信頼区間

効果量

- p値：標本の大きさが大きくなると、実質的な差がなかった場合でも小さな値にある

  - 標本の大きさによって変化しない「効果量」を併記することで、解釈に役立つ

効果量＝（群Aの平均－群Bの平均）／標準偏差

実際の差

効果量（式）

$$T = \frac{\bar{y}_A - \bar{y}_B}{s^* \sqrt{1/n_A + 1/n_B}}$$

$$= \boxed{\frac{\bar{y}_A - \bar{y}_B}{s^*}} \times \sqrt{\frac{n_A n_B}{n_A + n_B}}$$

効果量

比率に関する検定

比率に関する推定と検定の方法

- 正規近似を使う（大標本）
- そのまま2項分布（小標本）

比率＝
YesとNoを1
と0で
置き換えた
場合の平均

比率の検定

帰無仮説：$R = R_0$（R_0：数値）

対立仮説：$R \neq R_0$

棄却域は

$$\boxed{r < R_0 - 1.96\sqrt{\frac{R_0(1-R_0)}{n}}, \quad R_0 + 1.96\sqrt{\frac{R_0(1-R_0)}{n}} < r}$$

（有意水準５％、両側検定）

比率の検定（ある値との比較）

- 大標本の場合

帰無仮説：$R = R_0$ （R_0

対立仮説：$R \neq R_0$

$$\text{検定統計量}: t = \frac{r - R_0}{\sqrt{R_0(1 - R_0)/n}}$$

r：標本比率（データから計算）

n：データ数

この値と、
1.96（有意水準5%の場合）
を比較する

EXCELでp値を計算し、
定められた有意水準より小
さいかどうかをみる

例題 1

　ある工場の製品400個について検査したところ、不良品が18個あった。全製品における不良率が2%であるかどうかを危険率5%で検定せよ。

比率の検定１（例）

不良率：$q = 18/400 = 0.045$、帰無仮説：$p = 0.02$

$$0.02 - 1.96\sqrt{\frac{0.02(1-0.02)}{400}} = 0.006$$

$$0.02 + 1.96\sqrt{\frac{0.02(1-0.02)}{400}} = 0.034$$

なので、有意水準５％の棄却域は

$$q < 0.006, \quad 0.034 < q$$

$q = 0.045$ は上の棄却域に入っている。

よって、帰無仮説 $p = 0.02$ は棄却される。

比率の検定２（例）

不良率：$p = 18/400 = 0.045$、帰無仮説：$p_0 = 0.02$

検定統計量：

$$t = \frac{0.045 - 0.02}{\sqrt{0.02(1-0.02)/400}} = 3.5714$$

検定統計量の絶対値3.5714と、
有意水準5%の基準点1.96とを比較して、
検定統計量の絶対値の値の方が大きいので、
帰無仮説を棄却する。

p値:0.000355 （p値＜0.01）

レポート3

- 正規分布に関する問題
- 平均に関する推定と検定
- 比率に関する推定と検定