

■本資料のご利用にあたって(詳細は「利用条件」をご覧ください)

本資料には、著作権の制限に応じて次のようなマークを付しています。
本資料をご利用する際には、その定めるところに従ってください。

***** : 著作権が第三者に帰属する著作物であり、利用にあたっては、この第三者より直接承諾を得る必要があります。

CC : 著作権が第三者に帰属する第三者の著作物であるが、クリエイティブ・コモンズのライセンスのもとで利用できます。

 : パブリックドメインであり、著作権の制限なく利用できます。

なし : 上記のマークが付されていない場合は、著作権が東京大学及び東京大学の教員等に帰属します。
無償で、非営利的かつ教育的な目的に限って、次の形で利用することを許諾します。

- I 複製及び複製物の頒布、譲渡、貸与
- II 上映
- III インターネット配信等の公衆送信
- IV 翻訳、編集、その他の変更
- V 本資料をもとに作成された二次的著作物についての I からIV

ご利用にあたっては、次のどちらかのクレジットを明記してください。

東京大学 UTokyo OCW 学術俯瞰講義
Copyright 2014, 竹村彰通

The University of Tokyo / UTokyo OCW The Global Focus on Knowledge Lecture Series
Copyright 2014, Akimichi Takemura

統計学の歴史：確率・統計の歴史と考え方

工学部計数工学科 竹村彰通

2014 年 5 月 29 日

項目

1. 確率と統計の起源
2. 20 世紀の確率論と統計学
3. 統計的手法の性格

確率と統計の起源

「統計」という用語の起源

- 統計学は英語では statistics とよばれる. State は国家を意味しており, 歴史的には社会や国家の現状を数量的に把握するための方法として発展してきた.
- 「統計」という言葉は英語の statistics の訳語として、明治年間を通じて次第に定着するようになった.
- 「統計」を冠した最初の政府組織は明治4年に大蔵省に置かれた「統計司」と「統計寮」.
- 当時他にも「政表」「表記」「表紀」「形勢」などが提案された.

- 「統計」は 1900 年代初頭に日本の統計学関係の書籍を通じて中国に伝わってそのまま根付き，今日では中国語としても使われている．

「確率」という用語の起源

- probability あるいは chance の訳語
- 昭和になるまで，蓋然，可信，概然，偶然率，公算，などのさまざまな訳語が使われていた．戦後はようやく確率で定着した．
- 中国では「概率」とよぶようになった．

統計学の起源: 社会経済的

- 国力の定量的な評価 (statistics) 税金, 軍事
- 生命表, 生命保険 (グラントの死亡表 (1662), ハレー彗星のハレー (1693))
- 損害保険 (難破事故や海賊, リスク共有)
- 政治算術 ペティ 1690 年
- ケトラー (1796–1874) の大量観察, 「平均人」
- 白衣の天使 ナイチンゲール (1820-1910) の衛生統計

国力の定量的な評価

- 古代エジプト：ピラミッド建設のための調査
- ローマ帝国：初代皇帝アウグストゥスの治世の頃に、人口や土地を調べる調査（Census）が行われた。今日、国勢調査のことを「人口センサス」と呼ぶのはその名残。
- 中国：紀元前 2300 年に人口調査（孔子の文献による）
- 16 世紀以降のヨーロッパでは、人口や貿易の統計が整備されるようになってきた。
- ドイツの「国勢学」：人口，土地面積，歳入歳出，などを記録し表式で表す。

生命表, 生命保険

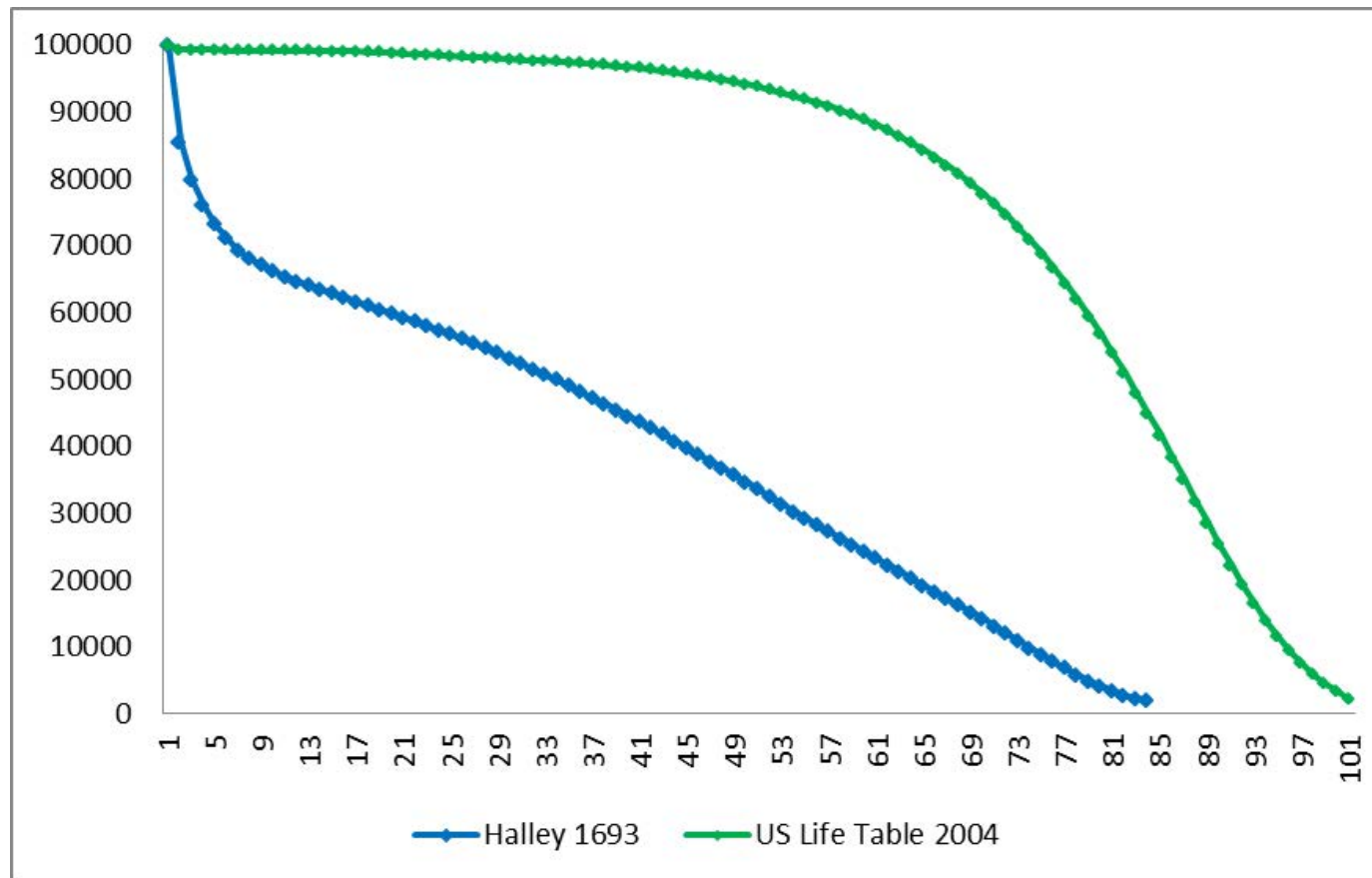
- グラントの死亡表 (1662): ロンドンの教会の過去帳に基づいて以下のような記述:
 - 出生者 100 人中, 6 才に達するまでに死亡する者 36 人, 以後 10 年を経過する毎に, 24, 15, 9, 6, 4, 3, 2, 1 とへり 86 年目には生存者が 0.
 - 出生時における男女比は 14:13.
- Halley の生命表: 年金制度や生命保険を意識して生命表を作成. Halley の表によって科学的に保険料率が算定されるようになった.

Halley の表

Age	Persons	Age	Persons	Age	Persons	Age	Persons
1	1000	23	579	45	397	67	172
2	855	24	573	46	387	68	162
3	798	25	567	47	377	69	152
4	760	26	560	48	367	70	142
5	732	27	553	49	357	71	131
6	710	28	546	50	346	72	120
7	692	29	539	51	335	73	109
8	680	30	531	52	324	74	98
9	670	31	523	53	313	75	88
10	661	32	515	54	302	76	78
11	653	33	507	55	292	77	68
12	646	34	499	56	282	78	58
13	640	35	490	57	272	79	49
14	634	36	481	58	262	80	41
15	628	37	472	59	252	81	34
16	622	38	463	60	242	82	28
17	616	39	454	61	232	83	23
18	610	40	445	62	222	84	20
19	604	41	436	63	212		
20	598	42	427	64	202	85-100	107
21	592	43	417	65	192		
22	586	44	407	66	182	Total	34000

James E. Cieccka (2008) Edmond Halley's Life Table and Its Uses. Journal of Legal Economics, vol.15 (no.1):65-74, p.67 Table 1に基づき作成

現在の生存率との比較



以下のデータを基に作成：

[Halley 1963] James E. Ciecka (2008) Edmond Halley's Life Table and Its Uses, *Journal of Legal Economics*, vol.15 (no.1):65-74, p.67 Table 1.

[US Life Table 2003] Elizabeth Arias (2007) United States Life Tables, 2004, *National Vital Statistics Reports*, vol.56 (no.9):1-40, p.8 Table 1.

損害保険

- 古代ギリシャ：嵐や海賊などが原因で、積荷を海に捨てた時、その損害を荷主と船主で負担する習慣が生まれた。
- 12～13 世紀頃の地中海沿岸：「冒険貸借」航海中に事故が起こった場合には、借入金を返さない。
- 日本でも江戸時代に初期から「朱印船貿易」で同様の制度があった。
- 1666 年のロンドンの大火災：火災保険のきっかけ

政治算術 ペティ

- 前書き

- 「私が採用した方法は... ただ比較級や最上級の言葉と英知的な議論とを用いる代わりに... 数と重さと尺度との用語で... 表現しようとする過程をとった。」
- 「特定人と変わりやすい精神，意見... などに左右される事項はこれを他の人々の考察に譲ったのである。」

- 外国との比較:
 - Chap. III. That France cannot by reason of Natural and Perpetual Impediments, be more powerful at Sea, than the English, or Hollanders.
 - Chap. IV. That the People, and Territories of the King of England, are Naturally near as considerable, for Wealth, and Strength, as those of France.
 - 「London では最下級の病院でも 16 人につき 2 人の死亡率であるのに, Paris では最上級の病院でも 15 人に 2 人の死亡率であるから, London の空気は Paris のそれよりも遥かに良いと認められる。」

ケトラーの「平均人」

- 確率論と統計学の融合のはじまり
- 社会現象の大量観察による「真理」の発見
- 「平均人」：平均身長，体重，知能
- 官庁統計の整備，統計の国際的な比較可能性の重視
- 1853：第一回国際統計会議の開催

例：ケトラーの犯罪統計

フランス刑事裁判報告による殺人件数

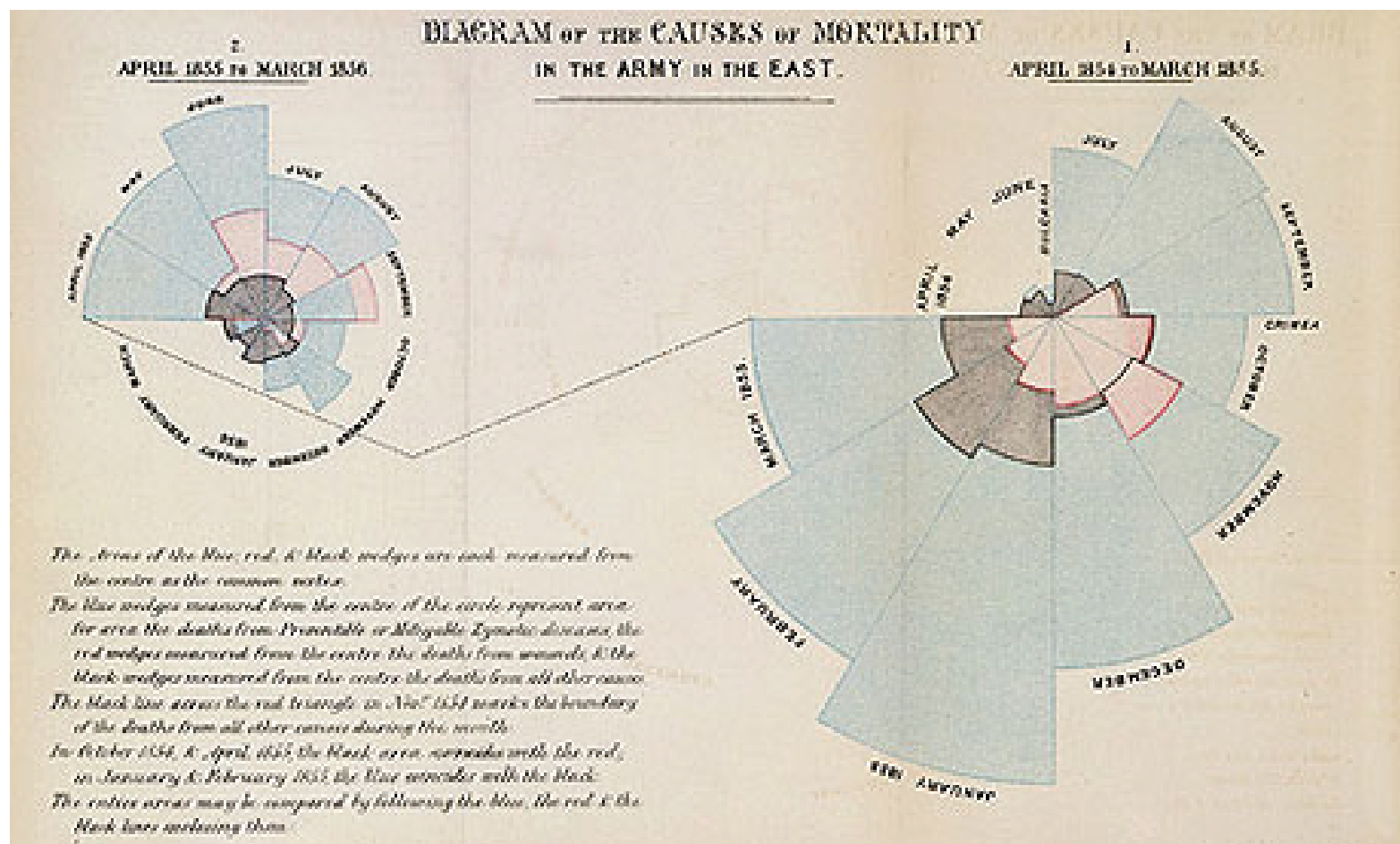
方法 \ 年別	1826年	1827年	1828年	1829年	1830年	1831年
総数	241	234	227	231	205	266
銃火器・短銃によるもの	56	64	60	61	57	88
剣・短刀によるもの	15	7	8	7	12	30
小刀によるもの	39	40	34	46	44	34
棒・杖等によるもの	23	28	31	24	12	21
石によるもの	20	20	21	21	11	9
打物・刺物・刃物	35	40	42	45	46	49
絞殺	2	5	2	2	2	4
投殺・溺殺	6	16	6	1	4	3
毒殺・撲殺	28	12	21	23	17	26
焼殺	0	1	0	1	0	0
不明	17	1	2	0	2	2

小杉肇『統計学史』恒星社厚生閣、1984年、p.188を基に作成

白衣の天使 ナイチンゲールの衛生統計

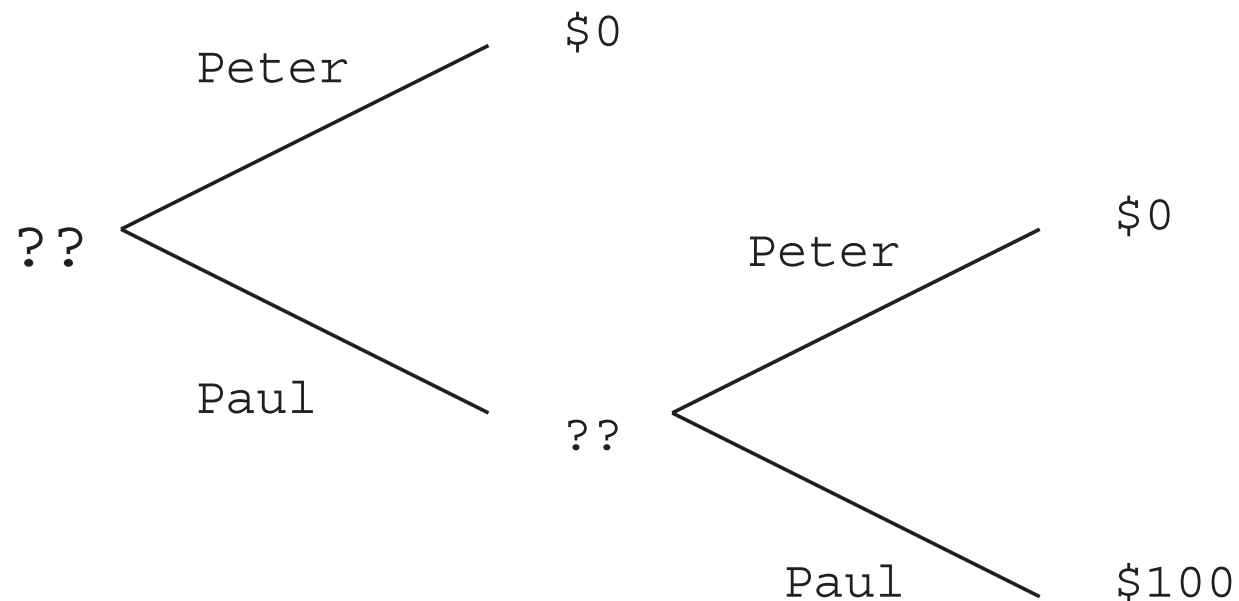
- イギリス軍の戦死者・傷病者に関する膨大なデータを分析
- 戦闘で受けた傷そのものではなく、その後の治療や病院の衛生状態が十分でないことが原因で死亡したことを明らかにした。
- 統計グラフを工夫して、議会でわかりやすく訴えた。

白衣の天使 ナイチンゲールの衛生統計

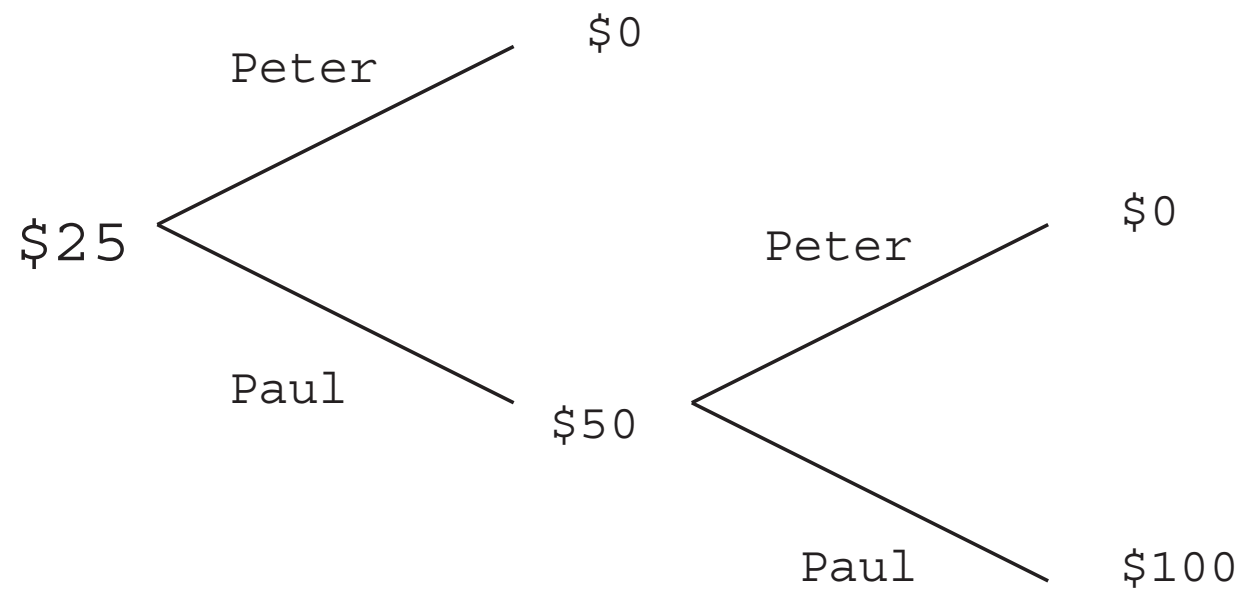


確率の起源

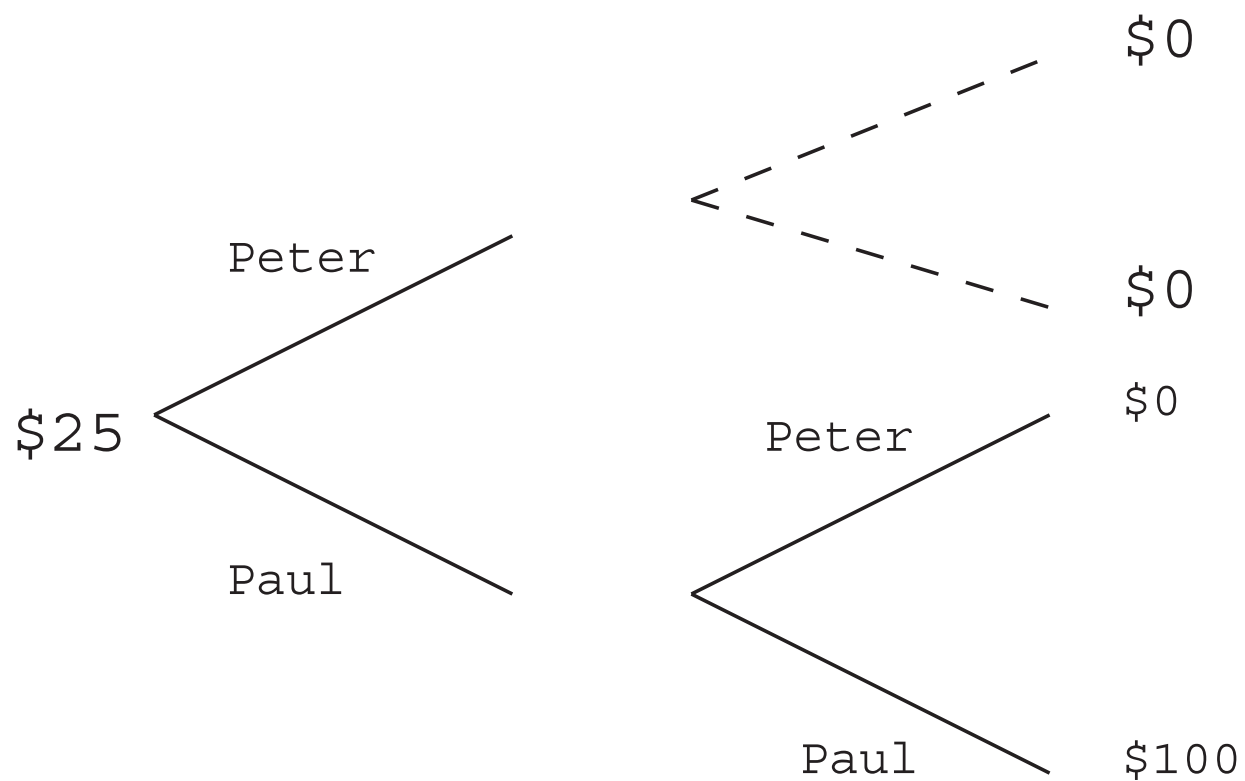
- テーブルゲーム (賭け) と確率論
フェルマーとパスカルの往復書簡 1654 「ゲームを途中でやめた時に, 賭け金をどう配分するか」



パスカルの答 (後ろ向き帰納法):



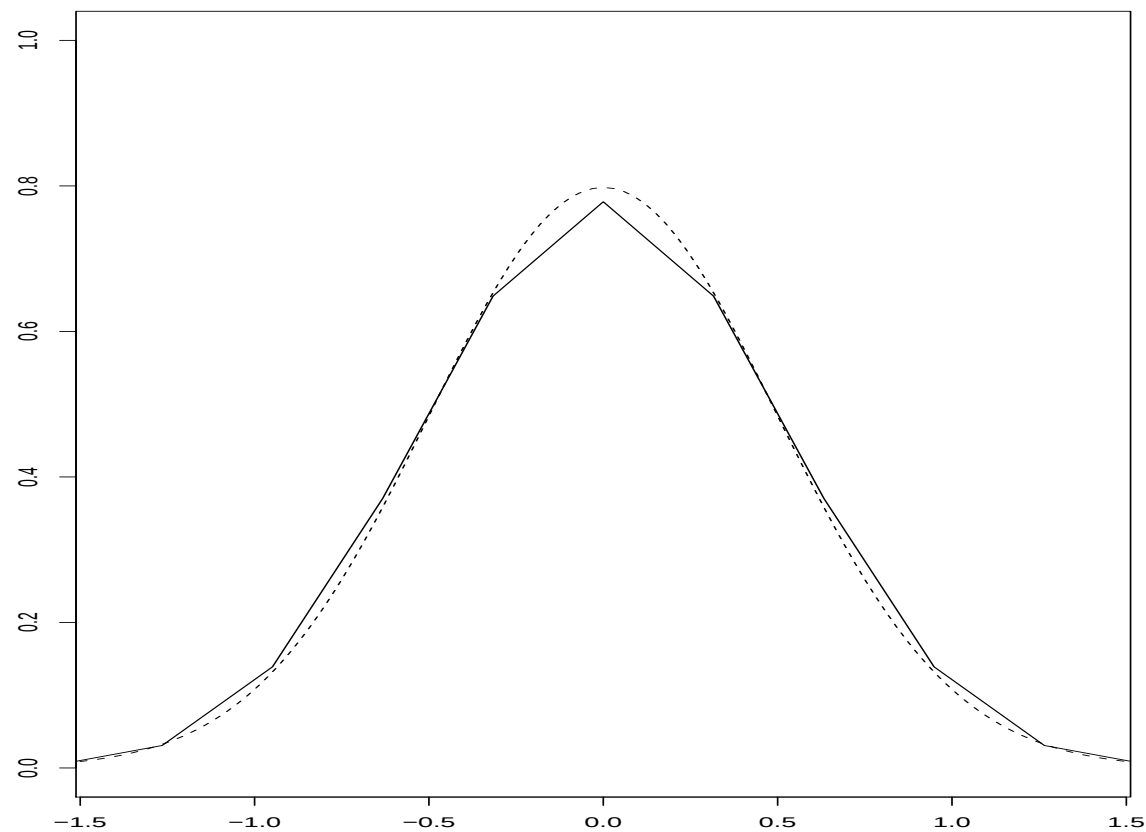
フェルマーの答 (同様に確からしい場合に分ける):



- 大数法則: ヤコブ・ベルヌーイ (“Ars Conjectandi” 1703)
- 中心極限定理: ド・モアブル Moivre 1733 (長く忘れられた), ラプラス 1812

ここまでは, 遊びの感じが強い. フランスの偉い数学者の趣味?

2項係数 ${}_{10}C_k$, $0 \leq k \leq 10$, (1,10,45,120,210,252,...)
と正規分布の密度関数 ($\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$)



- 天体観測にともなう誤差解析 – 最小二乗法,
正規分布 (ルジャンドル, ガウス)
- 遺伝法則と生物測定 (ゴルトン, カール・ピアソン)
相関, 回帰分析
遺伝が本質的に確率的だと理解されたことが大きい.

進化は? 免疫系は? (量子力学は?)

- メンデル (1865) はしばらく忘れられた.
- 1900 年に 3 人の別々の研究者 (ド・フリース, コレンス, チェルマック) によって再発見され、同じ雑誌に前後して掲載.
- メンデルのデータは「きれいすぎる」(3:1 等の比率) ことが、後に指摘されている.
- 以下は「工学のためのデータサイエンス入門」間瀬ほか、数理工学社、より:

実験者(年代)	黄個体数	緑個体数	計	比	p値
メンデル(1866)	6022	2001	8023	3.0095:1	0.9076
コレンス(1900)	1394	453	1847	3.0773:1	0.648
チェルマク(1900)	3580	1190	4770	3.0084:1	0.9467
ハースト(1904)	1310	445	1755	2.9438:1	0.2496
ベーストン(1905)	11903	3903	15806	3.0497:1	0.3779
ロック(1905)	1438	514	1952	2.7977:1	0.1742
ダヒシャイヤ(1909)	109060	36186	145246	3.0139:1	0.4488
ウィンゲ(1924)	19195	6553	25748	2.9292:1	0.095

間瀬茂・神保雅一・鎌倉稔成・金藤浩司『工学のためのデータサイエンス入門』
数理工学社、2004年、p.96を基に作成

20世紀の確率論と統計学

測度論的確率論の成立: コルモゴロフ

- (19世紀ロシアの確率論, カントールの集合論など)
- ルベーク積分 (1902)
- ブラウン運動 Einstein (1905)
- Kolmogorov (1933)

確率論を公理的に構成. 確率の意味を問わない.

数学の一分野として成立. 極限操作等の操作性の保証.

- 公理的ではあるが, やはり哲学がある. 特にランダム化が根源事象に集約されていると見ている.

「統計学」の成立: イギリスの統計学

- カール・ピアソン 「科学の文法」としての統計学
- R.A. フィッシャー (1890-1962): 数理統計学の創始者
大標本から小標本へ, 推測統計学,
実験計画法, 最尤法
- E.S.Pearson, J.Neyman 等

20 世紀前半の応用分野: 生物, 農業 (品種改良), 一部心理学でも

戦後の統計学

- アメリカ中心
- 「数理統計学」の成立 (Wald の統計的決定理論等)
 - 数学の形式化の流れ
 - 独自の理論体系と内的発展の重視

応用分野の拡がり

- 計量経済学（大規模計量モデルの功罪）
- 心理学，社会学（アメリカ的な数量的な研究の重視）
[我が国で言う「文系」とはやや異なる]
- 医学統計
大規模臨床試験，新薬の認可
副作用，Evidence Based Medicine

- 実験計画法の工業への応用
 - － 統計的品質管理
 - － 日本からの貢献も大きい (因子の性質の区別等)
 - － 「田口メソッド」 (田口玄一)
品質のばらつきに影響を与える要因の分析の重視
ロバスト設計
 - － 製造業・大量生産に有効な面

- 金融工学，数理ファイナンス（1980年代より，主に確率論）
 - － ブラック・ショールズ式などが注目を集めた
 - － 株価変動などの金融データを確率過程としてモデル化
 - － 伊藤積分など確率解析が重要な役割
 - － 派生証券の価格づけ理論は規範的な理論
 - － 理論と現実との乖離　特にリーマン・ショック（2008/9）以降
 - － 統計的なデータ解析もさかんにおこなわれている

- ゲーム理論における混合戦略
 - フォン・ノイマン及びモルゲンシュテルンによる古典的な二人ゼロ和ゲームでは，ゲームの相手が賢いと考ええると，自然に確率化戦略が最適な戦略となる．
 - じゃんけんの例
 - 確率の意味の理解の上でも重要（確率は「思考の単純化」？）

確率モデルと統計手法の変化

背景として，計算機の発達により，実際に複雑な計算ができるようになったという事実がある．

1. 統計モデルの大規模化

データ自身が大量に得られる．これにともない統計モデル (データを説明する確率分布) も大規模化してきた．大規模モデルの計算にともなう面が手法的には重要となってきた．

2. モデル選択の方法 (最尤法ベース)

- モデルの説明力 (データへのフィット) とモデルの簡潔さのバランス
- 複雑なモデルは手もとのデータへのフィットは高いが, 新たなデータの予測が不安定となる場合がある (overfitting, 「過学習」)
- AIC (赤池情報量規準) などの手法によるモデル選択. 我が国の貢献が大きい.

3. ベイズ法

- システムのパラメータに関する不確実性や，分析者の主観的な情報などをすべて確率変数としてモデル化する (事前分布).
- システムの未知変数は単に観測されていない確率変数であると考える.
- 統計的推論は，データに基づいて事前分布を事後分布でおきかえる．機械的な操作となる．概念的にはすべて確率計算のみとなる．

- 1960年代には、客観確率と主観確率の違いなどに関する哲学的な議論があったが、最近では「手法」と割り切って考える.
- 機械的な操作で統計的推論をおこなえるという点が、計算機との相性がよい.
- 観測されない「潜在変数」の扱いも比較的容易で、扱えるモデルのクラスが広い.

4. モンテカルロ手法

- シミュレーションの一種であるが，もっぱら乱数を用いた方法.
- 確率計算にともなう積分などの計算を，乱数を用いて近似的におこなう.
- 近似の誤差はあるが単純で汎用性がある.
- 「マルコフ連鎖モンテカルロ法」の導入により，乱数発生自体を近似的におこなうことが可能となった.

5. マルコフ連鎖モンテカルロ法 (MCMC, Markov chain Monte Carlo)

- 確率モデルに従う乱数を発生させたいが、モデルの従う乱数そのものが発生しにくい場合がある.
- 単純化して言うと、モデルを解く際の収束計算のステップを、乱数をもとにおこなうことにより、近似的な乱数を発生することができる.
- モンテカルロ法の適用範囲が大きく広がった.

統計的方法の性格

「統計」の性格

- 均一性を持つ母集団の想定し，母集団におけるバラツキの存在を前提
- 生命保険の例
 - － 性別，年齢等で分ければ均一の集団と考えられ，そこでリスクをプールする．
 - － 「均一性」を信じるには情報が少ないほうがいいこともある．

- 個人の遺伝子情報などがとれて情報が多くなりすぎると、リスクをプールできなくなる可能性がある。
- 保険ではなくて、社会保障でカバーするのか。
- (じゃんけんに勝てる機械，サイコロは乱数?)
- 一定の重要な要因を所与とすると，それ以上は純粋な確率的な変動であり説明できないとわりきって考える。
- 確率的な変動は多数回反復可能であり，「リスクをプールする」ことにより，平均的な性能が支配的な状況を設定する。

- 薬の例

- 平均的に効く薬がいいのか,「特効薬」がいいのか
- とりあえずの答: (劇的な効果はなくても) どのような条件の人にも, 平均的に効く薬がいい.
- 「交互作用」のないものがよい.
- 「ロバスト性」,「副作用」
- 「特効薬」といっても, 一定の大きさの母集団は仮定されている.

統計的方法の性格

- 統計モデル = 固定効果 + 誤差
- 誤差はブラックボックス
- 固定効果は未知母数を含む
- 誤差の分布について平均的な性能で評価する

確率計算万能主義

- 定数も分散の小さい確率変数
- 未知母数の確率変数 (ベイズ法)
- すべては確率計算で済む.
- 大量のデータに語らせる. 意味は問わない.

参考文献と参考にしたサイト

1. James E. Ciecka (2008). “Edmond Halley’s Life Table and Its Uses.” *Journal of Legal Economics*, 15(1): pp. 65-74.
2. キース・デブリン (原啓介訳) (2010) 『世界を変えた手紙 – パスカル、フェルマーと〈確率〉の誕生』, 岩波書店.
3. 小杉 肇 (1984) 『統計学史』, 恒星社厚生閣.
4. 間瀬茂, 鎌倉稔成, 金藤浩司, 神保 雅一 (2004) 『工学のためのデータサイエンス入門』 数理工学社

5. 中塚利直 (2010) 『応用のための確率論入門』, 岩波書店.
6. 総務省統計局統計学習サイト統計学習の指導のために
(先生向け)
<http://www.stat.go.jp/teacher/index.htm>
7. Stephen M. Stigler (1986). *The History of Statistics*, Harvard University Press.