# 第62回　パートナー会　議事録

日時　2016月10月23日(日)　1時～5時
場所　CIS会議室

1)講師　久米 健次 様
　課題　　ベイズ統計


会議風景

2)第63回パートナー会議
　2017 年1月15(日)
　生駒 篤一 様

# ベイズ統計

## Bayesian statistics

at CIS Oct. 22, 2016

## ビッグデータ、IoT、AI、機械学習が流行。

センサー、ネットで膨大なデータがとれるようになった。
⇒ その中から役に立つ兆候や情報を引き出せるか。

今に始まったわけではない
データマイニング
ニューラルネット（機械学習は第３次ブーム）

何度もブームは来ては去り・・・の繰り返し。

基盤技術・・・確率・統計学

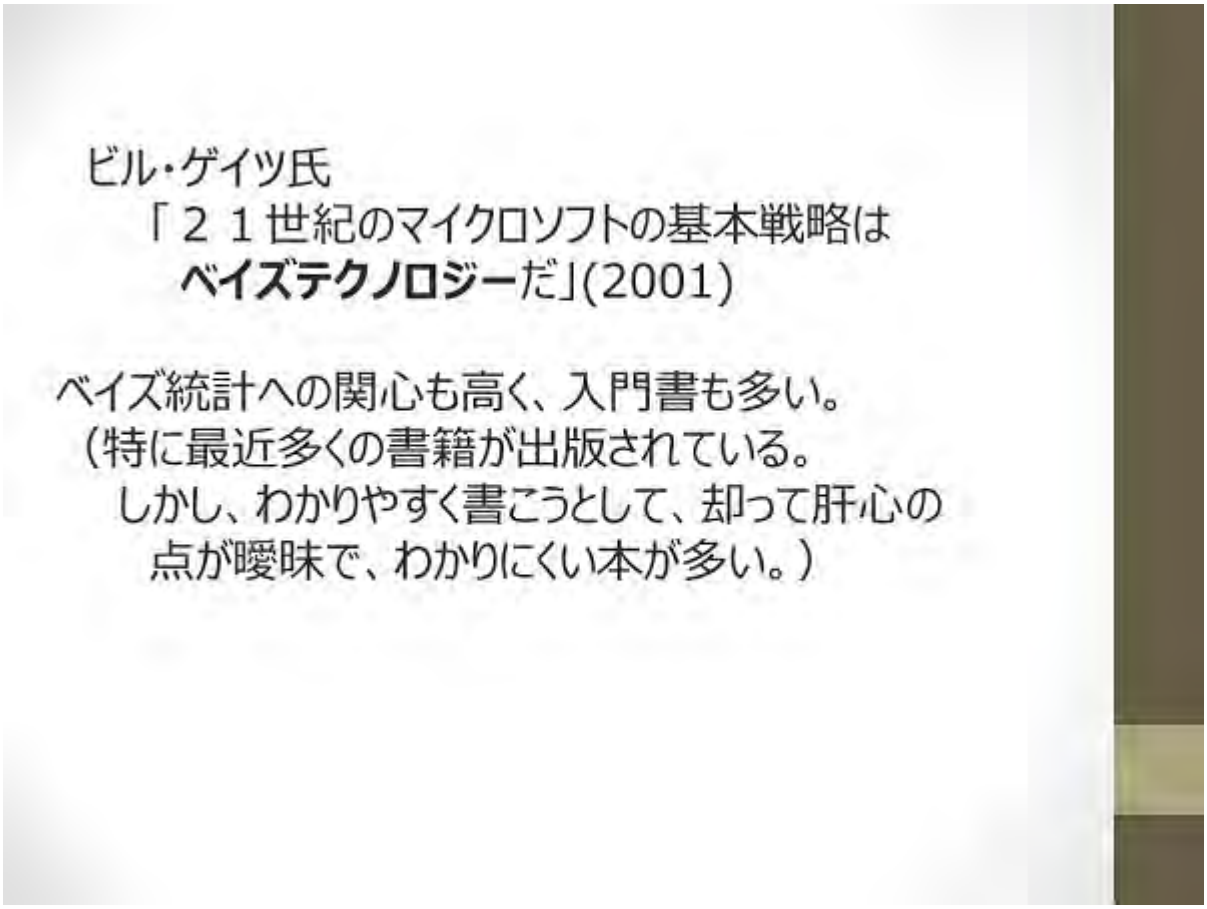
部長：田中君、わが社も
ビッグデータを
活用しようじゃないか。

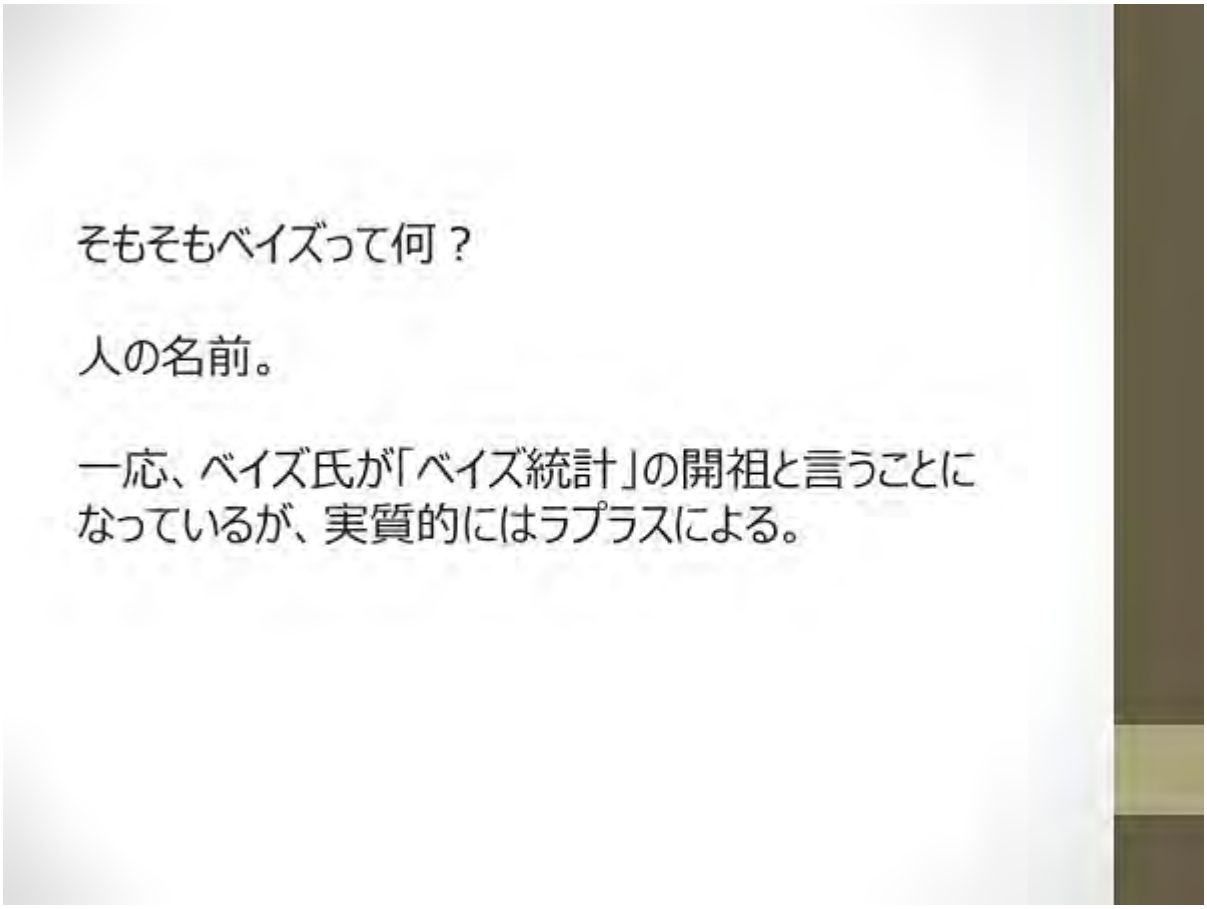

田中君：はっ。すぐさま検討します。

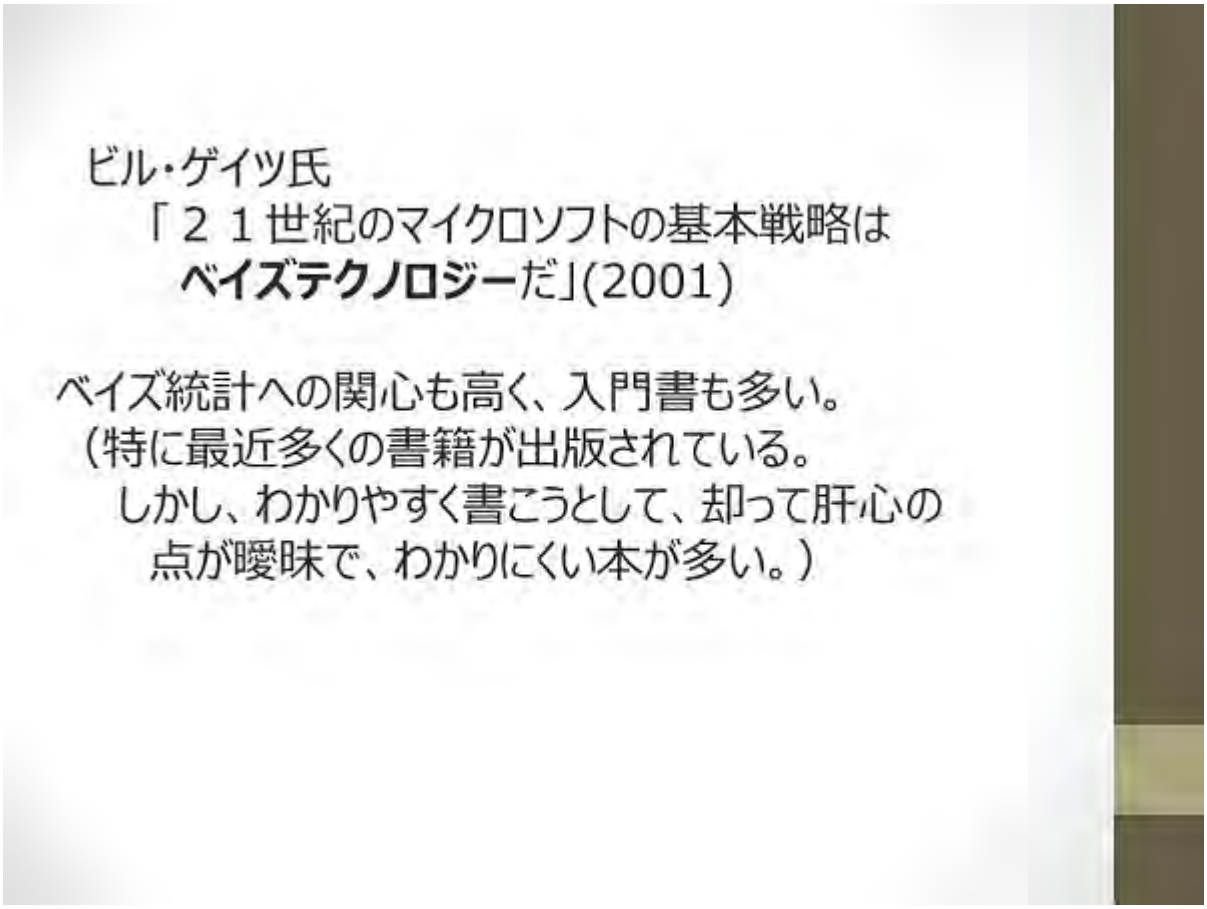
田中君は本屋に行って、
ビッグデータ関係の本を購入・・ドロナワ勉強。

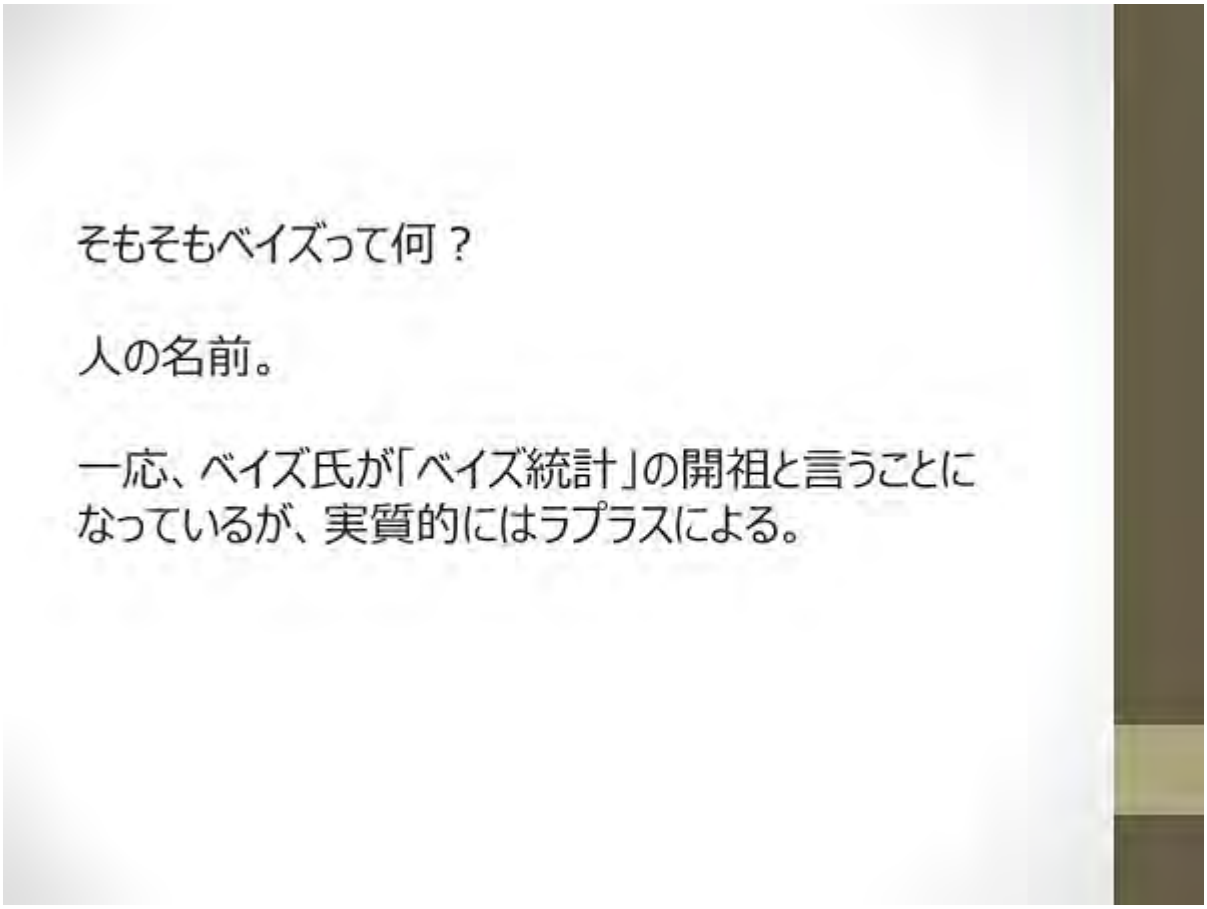
**本屋では「確率・統計＋ビジネス」本が平積み。**

- **統計学が最強の学問である**
  －データ社会を生き抜くための武器と教養
- **「人生成功」の統計学**
  －自己啓発の名著50冊に共通する8つの成功法則
- **統計学が世界を予測する**
  －図解　ビッグデータを制する者はビジネスを制す！
- **統計学でわかるビッグデータ**
- **ポスト・ビッグデータと統計学の時代**
- **ビッグデータ時代のマーケティング　ベイジアンモデリングの活用**
- **確率思考の戦略論 USJでも実証された数学マーケティングの力**
- **確率と統計によるネットワークビジネス成功への方程式**
- **図解・ベイズ統計「超」入門**
  －あいまいなデータから未来を予測する技術
- ・・・・・


ビル・ゲイツ氏
「２１世紀のマイクロソフトの基本戦略は
ベイズテクノロジーだ」(2001)
ベイズ統計への関心も高く、入門書も多い。
（特に最近多くの書籍が出版されている。
しかし、わかりやすく書こうとして、却って肝心の
点が曖昧で、わかりにくい本が多い。）


そもそもベイズって何？
人の名前。
一応、ベイズ氏が「ベイズ統計」の開祖と言うことに
なっているが、実質的にはラプラスによる。


ビル・ゲイツ氏
「２１世紀のマイクロソフトの基本戦略は
ベイズテクノロジーだ」(2001)
ベイズ統計への関心も高く、入門書も多い。
（特に最近多くの書籍が出版されている。
しかし、わかりやすく書こうとして、却って肝心の
点が曖昧で、わかりにくい本が多い。）


そもそもベイズって何？
人の名前。
一応、ベイズ氏が「ベイズ統計」の開祖と言うことに
なっているが、実質的にはラプラスによる。

Thomas Bayes　　1702年 - 1761年4月17日
（トーマス・ベイズ）

1936年に出版された『生命保険の歴史』にあるベイズの肖像画。
これが実際にベイズを描いているかどうかどうかは疑わしい。

ピエール＝シモン・ラプラス（仏）
（Pierre-Simon Laplace, 1749年3月23日 - 1827年3月5日）

統計学の分野では、

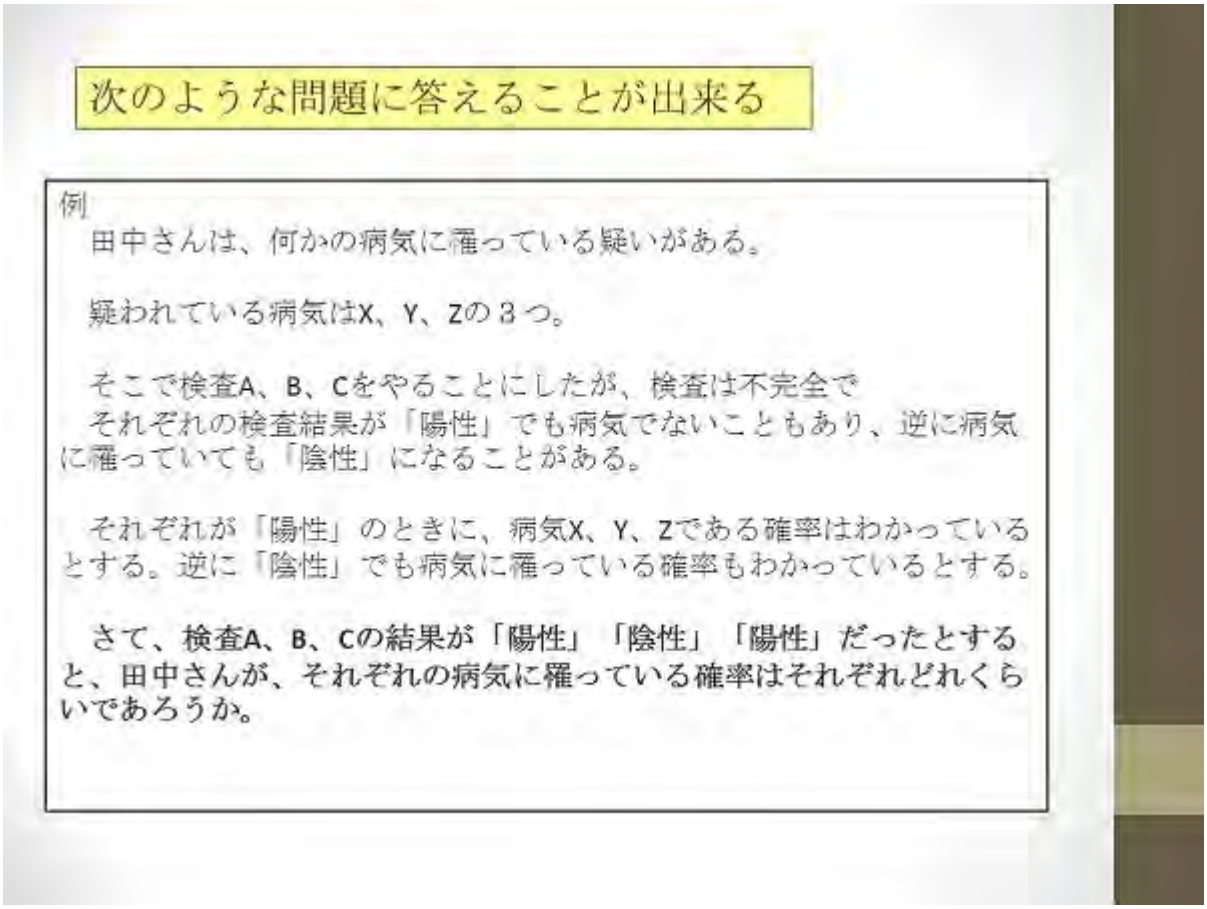
「古典統計学派（頻度主義）」と
「ベイズ派」

の150年を超える壮絶バトルがくりひろげられた。

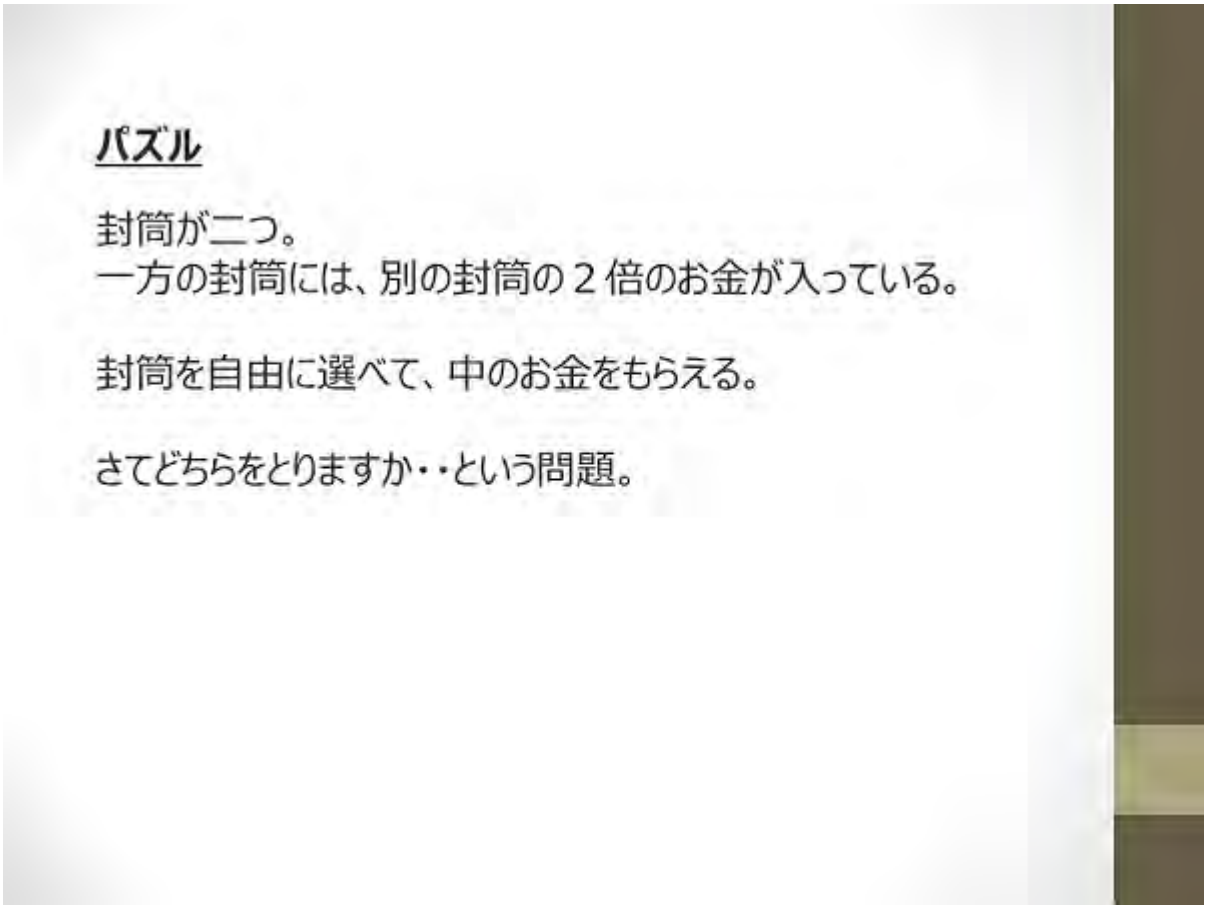

統計学の分野では、

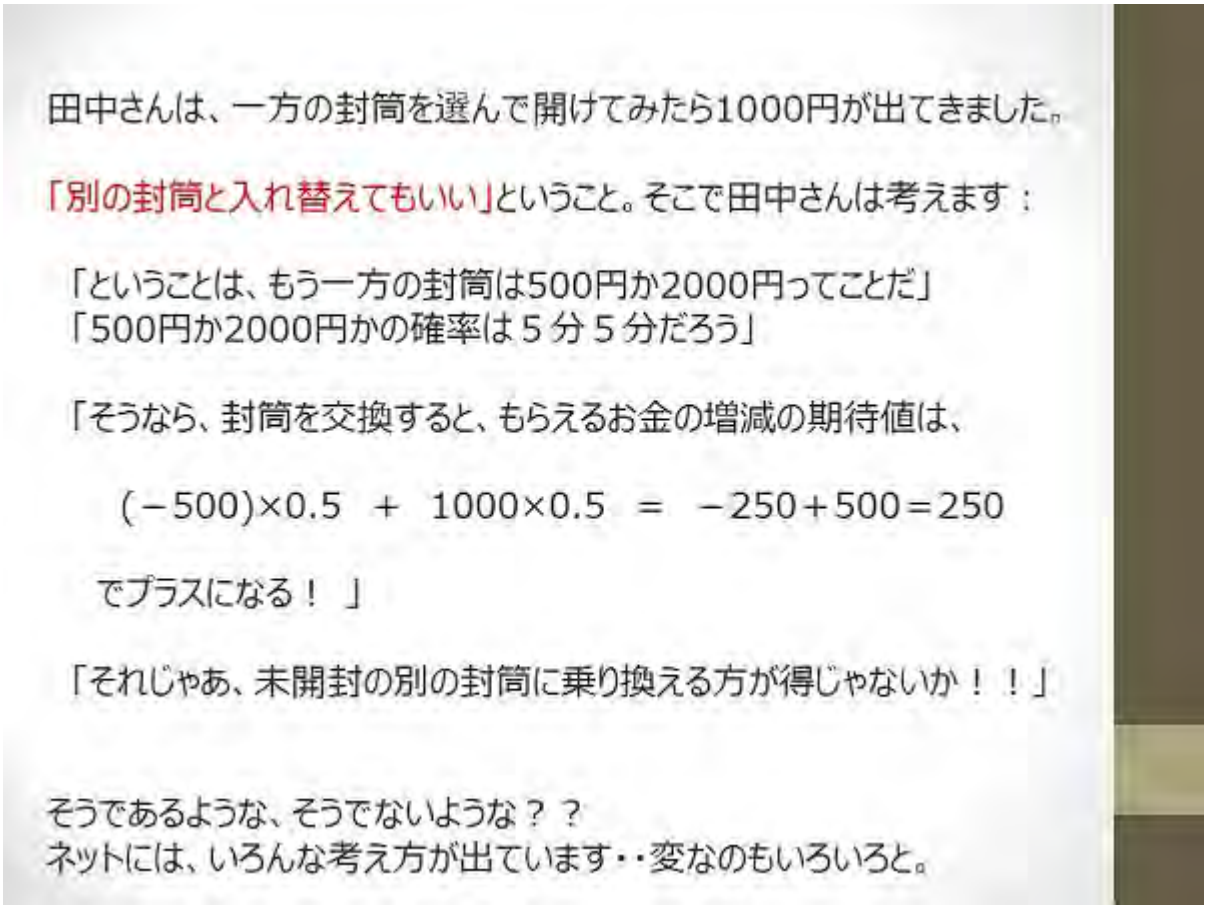
「古典統計学派（頻度主義）」と
「ベイズ派」

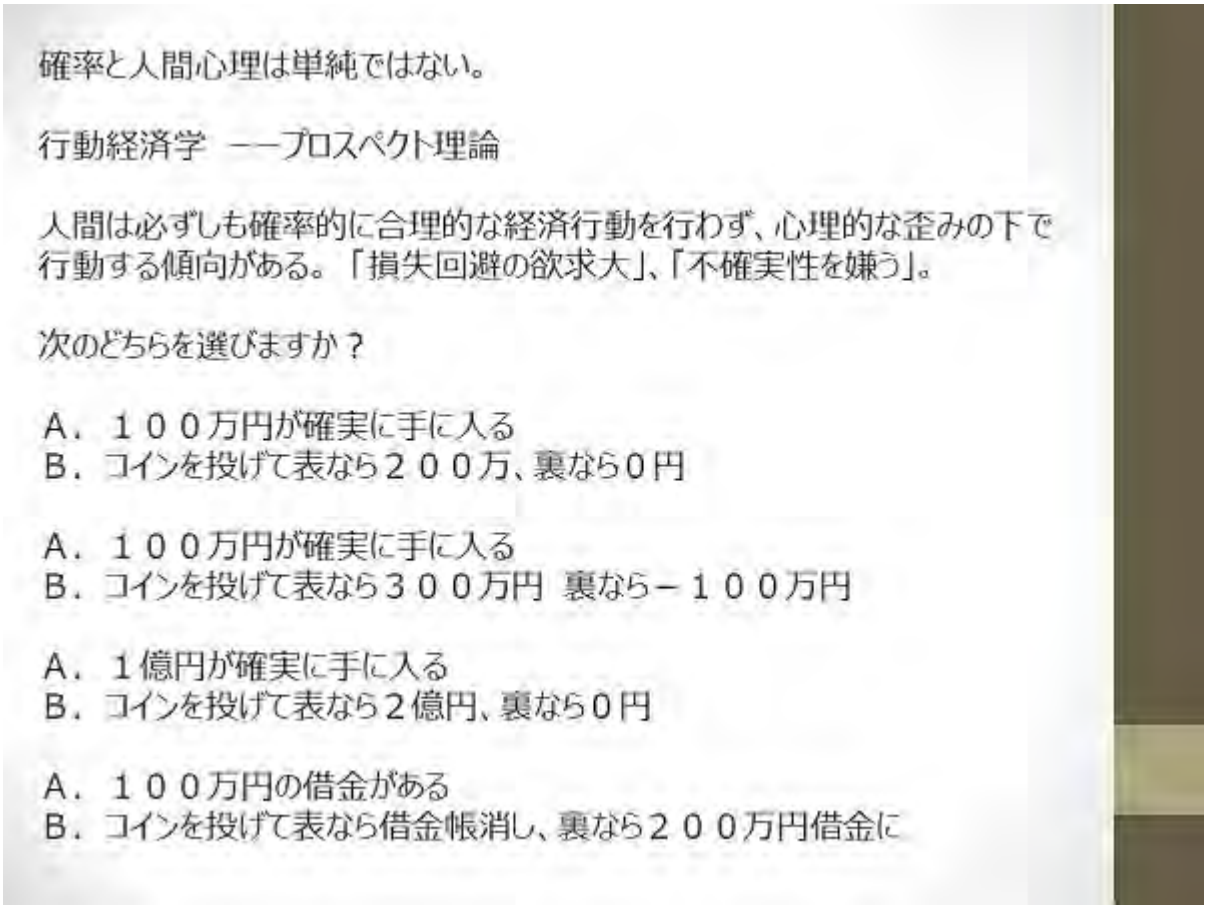
の150年を超える壮絶バトルがくりひろげられた。

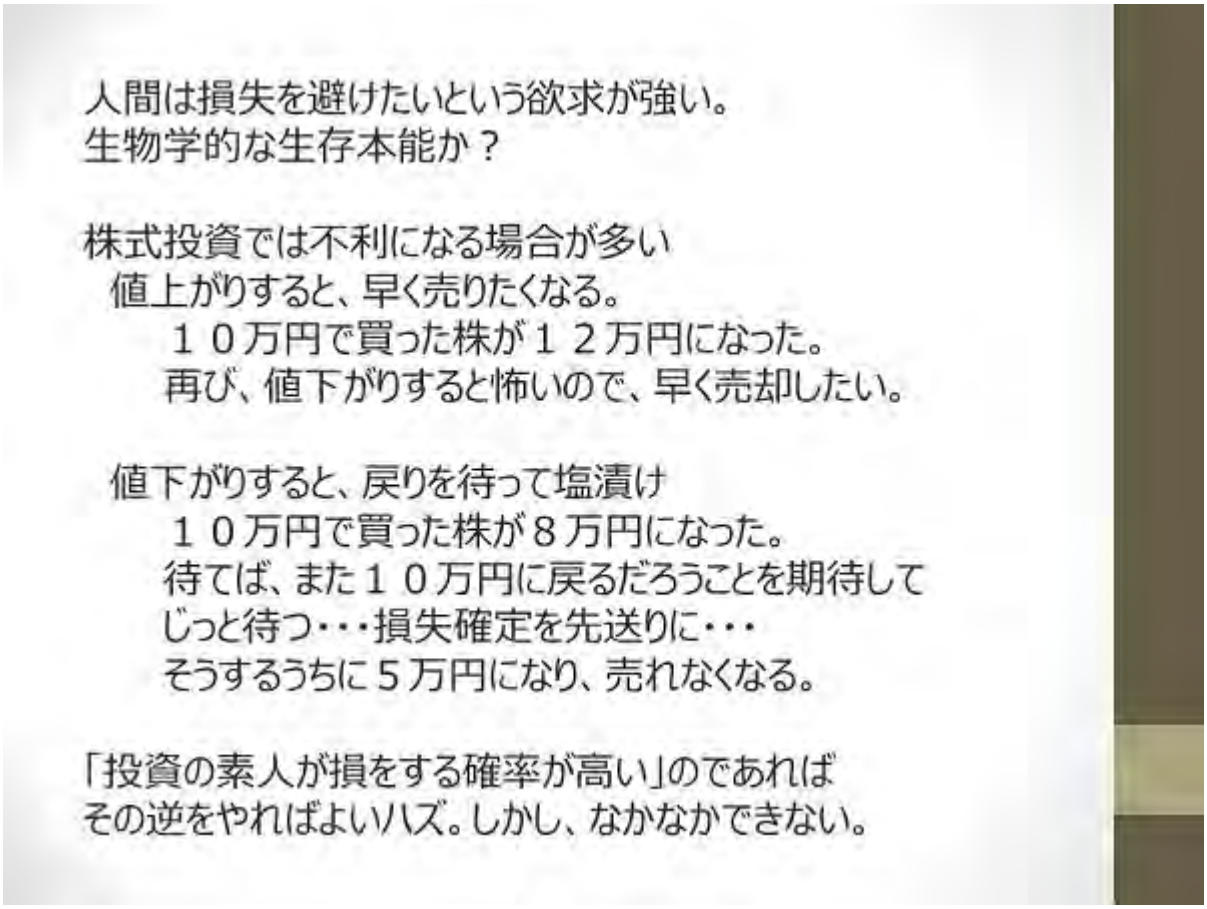

次のような問題に答えることが出来る
例
田中さんは、何かの病気に罹っている疑いがある。
疑われている病気はX、Y、Zの３つ。
そこで検査A、B、Cをやることにしたが、検査は不完全で
それぞれの検査結果が「陽性」でも病気でないこともあり、逆に病気に罹っていても「陰性」になることがある。
それぞれが「陽性」のときに、病気X、Y、Zである確率はわかっているとする。逆に「陰性」でも病気に罹っている確率もわかっているとする。
さて、検査A、B、Cの結果が「陽性」「陰性」「陽性」だったとすると、田中さんが、それぞれの病気に罹っている確率はそれぞれどれくらいであろうか。


ベイズの前に、確率のパズルを１つ。


パズル
封筒が二つ。
一方の封筒には、別の封筒の２倍のお金が入っている。
封筒を自由に選べて、中のお金をもらえる。
さてどちらをとりますか・・という問題。


田中さんは、一方の封筒を選んで開けてみたら1000円が出てきました。
「別の封筒と入れ替えてもいい」ということ。そこで田中さんは考えます：
「ということは、もう一方の封筒は500円か2000円ってことだ」
「500円か2000円かの確率は５分５分だろう」
「そうなら、封筒を交換すると、もらえるお金の増減の期待値は、
(−500)×0.5 ＋ 1000×0.5 ＝ −250＋500＝250
でプラスになる！ 」
「それじゃあ、未開封の別の封筒に乗り換える方が得じゃないか！！」
そうであるような、そうでないような？？
ネットには、いろんな考え方が出ています・・変なのもいろいろと。


確率と人間心理は単純ではない。
行動経済学　――プロスペクト理論
人間は必ずしも確率的に合理的な経済行動を行わず、心理的な歪みの下で
行動する傾向がある。「損失回避の欲求大」、「不確実性を嫌う」。
次のどちらを選びますか？
A．１００万円が確実に手に入る
B．コインを投げて表なら２００万、裏なら０円
A．１００万円が確実に手に入る
B．コインを投げて表なら３００万円　裏なら－１００万円
A．１億円が確実に手に入る
B．コインを投げて表なら２億円、裏なら０円
A．１００万円の借金がある
B．コインを投げて表なら借金帳消し、裏なら２００万円借金に


人間は損失を避けたいという欲求が強い。
生物学的な生存本能か？
株式投資では不利になる場合が多い
値上がりすると、早く売りたくなる。
１０万円で買った株が１２万円になった。
再び、値下がりすると怖いので、早く売却したい。
値下がりすると、戻りを待って塩漬け
１０万円で買った株が８万円になった。
待てば、また１０万円に戻るだろうことを期待して
じっと待つ・・・損失確定を先送りに・・・
そうするうちに５万円になり、売れなくなる。
「投資の素人が損をする確率が高い」のであれば
その逆をやればよいハズ。しかし、なかなかできない。

さて、話を戻して

・古典統計学とベイズ統計学は何が違うのか？

・何故、そんなにも長く激しい論争が続いたのか？

・現在の情勢はどうなのか？

・何に役に立っているのか？

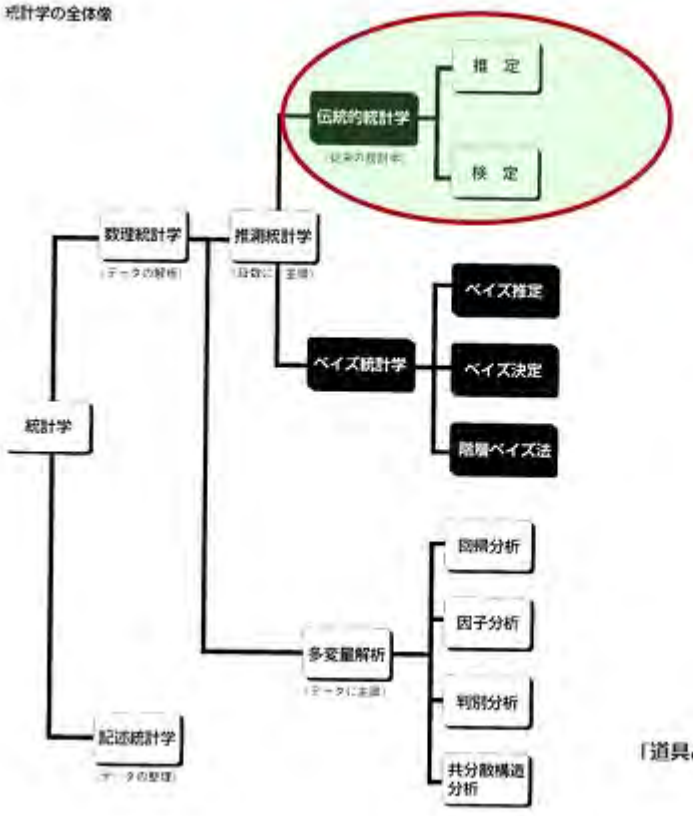
ベイズ：結果から原因を推定するのが得意

売り上げが伸びた→何故

病院の検査で陽性→何の病気

アラームが鳴った→どこの異常

糖尿病→何が主原因




「道具としてのベイズ統計」より

## 『異端の統計学ベイズ』（草思社:2011　訳2013）
## Sharon Bertsch McGrayne著より

### ベイズ統計の苦難の歴史

ベイズ統計を巡る科学者の戦いは、長く150年も続いた。
そこで論じられたのは、**「経験から学ぶ」**ということ。

つまり、**新たな情報が入った時に、初めの考え方を変えて、よりよい意見にどのように到達するか、と。**

ベイズの定理を受入れたに多くの人は、この定理の内なる論理に魅了された瞬間に、宗教的な啓示ともいうべき経験をする。

しかし、この定理を認めない人々にとっては、ベイズの法則は単なる主観性の暴走（勝手やり放題）でしかなかった。

18世紀中ごろ。

ベイズの定理を発見したのはアマチュア数学者
トーマス・ベイズ牧師（英）。

彼は自分が発見したことを積極的に売り込んだわけではなかった。当時、ベイズはさほど有名ではなく、ベイズの業績が今日に伝わっているのは、ひとえに彼の友人で編集者であったリチャード・プライスのおかげなのだ。

これは、本来はラプラスの法則と呼ぶべきものである。

フランスの極めて優れた数学者であったラプラスは、1774年に独力でこの法則を発見した。ラプラスが発見し、整備したこの法則を今でもベイズの法則と呼ぶのは慣習に過ぎない。

ラプラスの死後、厳密で客観的な理論を求める統計学者たちは、「こんなものは、主観的で役に立たぬ」と言って、ラプラスの手法を切って捨て、葬り去った。

だが、その一方で実際的な現場の問題と格闘している人々は、この法則をより所にして、現実世界の緊急事態に対処していった。

やがて、この法則は第二次世界大戦で華々しい成功を収めることとなる。

アラン・チューリングがこの法則を発展させて、ドイツ海軍の暗号を解いた。

ロシアのコルモゴロフやニューヨークにいたクロード・シャノンなどの一流の数学者たちが、戦時下での意思決定に役立てるべく「ベイズの法則」を見直し始めた。

行方不明の水爆や、潜水艦の位置の探査、原子力発電所の安全性、スペースシャトルの事故の予測、喫煙ががんを引き起こすこと、コレステロール値が高いと心臓発作が起きやすいこと・・・などを示すのにベイズ統計が使われたのである。

さて、冷静なはずの科学者が、ことベイズの法則に関しては極めて過剰なまでの反応と、激しいやりとりを行ったのはなぜだろうか。

答えは至極簡単。 ベイズの法則の核となるものが、科学者の心に深く根差した**「近代科学には正確さと客観性が要求される」**という心情に反していたからだ・・・主観が入ってもよいとした。

このような根本哲学の不一致があった。
頻度主義者によるベイズ派の圧倒的包囲網に向かって、このやり方が正しいことを認めさせ、受け入れさせようとする少数の擁護者たちの戦い、血塗られた苦闘の歴史が繰り広げられることになる。

第１章　１７４０年代～１７６４年
　　発見者ベイズ自身に見捨てられた大発見

第２章　１７７３年～１８２７年
　　「ベイズの法則」を完成させた男　ラプラス

第３章　１８２７年～１９３０年
　　ベイズの法則への激しい批判
　　反ベイズ３人組
　　「フィッシャー、エゴン・ピアソン、ネイマン」
　　（頻度主義統計学の完成者）による猛攻
　　ベイズ派　ジェフリーズ（地球物理学者）の孤立

第４章　１９３９年～１９５４年
　　ベイズ、戦争の英雄となる（日陰で）

第５章　１９４５年～１９５０年代
ベイズ、再び忌むべき存在となる

第６章　１９５０～１９６０年代前半
ベイズ主義の反撃
（保険数理士の世界からはじまった反撃）

第７章　１９５０～１９６０年代
ベイズを体系化し、哲学した３人
グッド、サヴェッジ、リンドレー

第８章　１９５０～１９７９年
ベイズ肺がんの原因を発見する

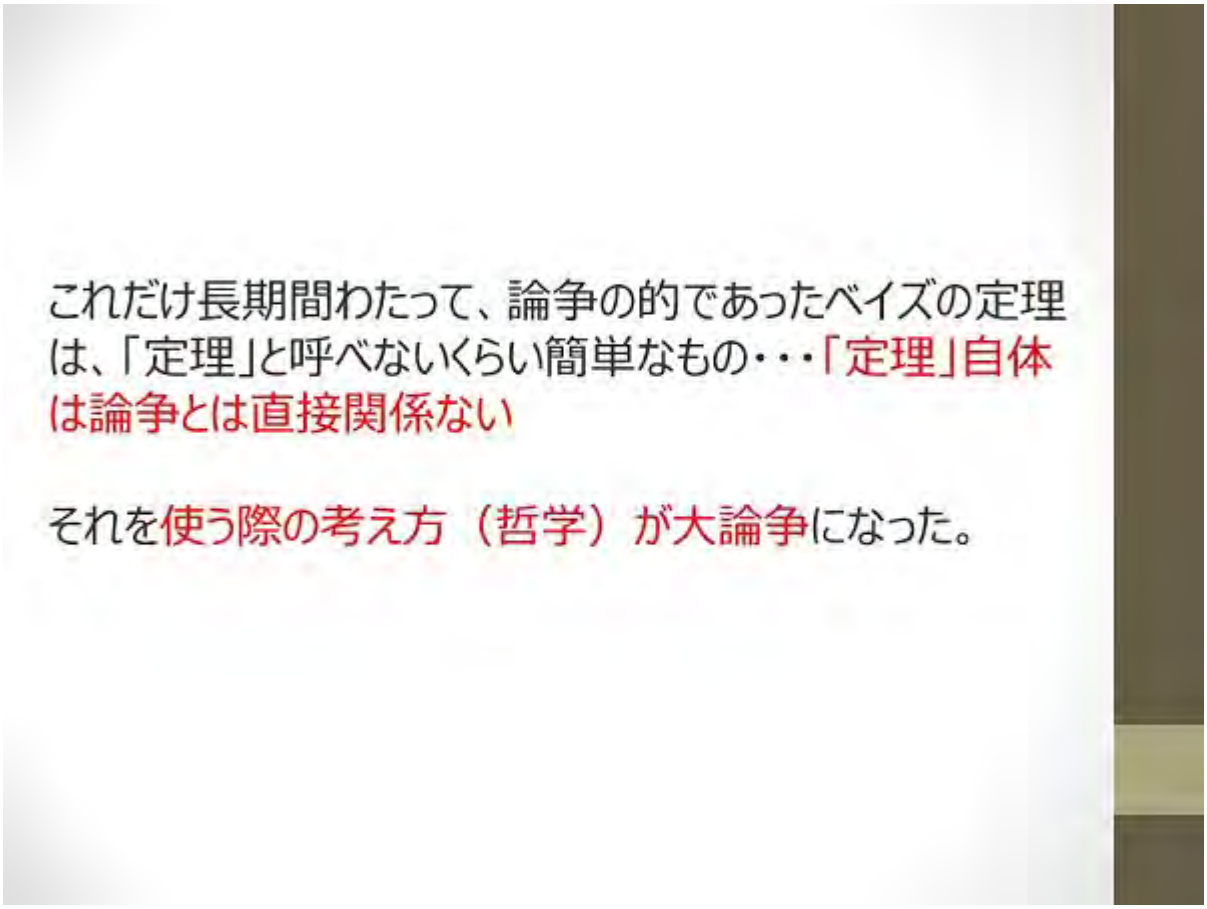
第９章　１９５７～１９５８年
核配備が進行する中、事故の確率をベイズ統計で算出

第１０章　１９５７～１９６０年代半ば
ベイズ派の巻き返しと、ネイマンらの頻度主義者の論争最高潮に

第１１章　１９５７～１９６５年
意思決定にベイズを使う

第１２章　１９５５年～１９６４年
単語の用法をベイズで分析し、論文の著者を推定


第１３章　１９６０～１９８０年
諜報業界の大物テューキが大統領選挙予測にベイズを使って成功も、その事実の公表は禁止に。
第１４章　１９７０～１９８１年
ベイズ派停滞も、スリーマイル島原発事故を予見
第１５章　１９６６～１９７６年
ベイズで海に消えた水爆や潜水艦を探す
第１６章　１９８０～２００８年
決定的なブレークスルー（多重積分　MCMC法の活用）
第１７章　現在～未来
世界を変えつつあるベイズ統計学
ベイズは受け入れられ、活用され、論争は沈静化した。


これだけ長期間わたって、論争の的であったベイズの定理は、「定理」と呼べないくらい簡単なもの・・・「定理」自体は論争とは直接関係ない
それを使う際の考え方（哲学）が大論争になった。

# ベイズの定理とは？

| | タバコを吸う | タバコを吸わない |
|---|---|---|
| 1000万円以上 | １００万人 | ２００万人 |
| 1000万円以下 | ３００万人 | ４００万人 |

全部で１０００万人。
年収が1000万円以上で、かつ、タバコを吸う人の確率（割合）は、
表を見れば、あきらかに　１００／１０００＝１／１０

このことを少し違う風に見てみる。

| | タバコを吸う | タバコを吸わない |
|---|---|---|
| 1000万円以上 | １００万人 | ２００万人 |
| 1000万円以下 | ３００万人 | ４００万人 |

P（1000万以上）＝３／１０
P（タバコを吸う｜1000万以上）＝１／３
⇒ P（タバコを吸う｜1000万以上）P（1000万以上）
＝１／１０

| | タバコを吸う | タバコを吸わない |
|---|---|---|
| 1000万円以上 | １００万人 | ２００万人 |
| 1000万円以下 | ３００万人 | ４００万人 |

P（タバコを吸う）＝４／１０
P（1000万以上｜タバコを吸う）＝１／４
⇒ P（1000万以上｜タバコを吸う）P（タバコを吸う）
＝１／１０

**条件付き確率に関する関係式**

⇒ **P（B｜A）P（A）＝P（A｜B）P（B）**

例

・インフルエンザが流行していて、人口の**３％が罹患**している。

・検査薬は不完全で

| | |
|---|---|
| インフルエンザにかかっていて陽性の確率 | 0.98 |
| インフルエンザにかかっていて陰性の確率 | 0.02 |
| インフルエンザにかかってなくて陽性の確率 | 0.05 |
| インフルエンザにかかってなくて陰性の確率 | 0.95 |

**さて、ある人が検査で陽性でした。**
**インフルエンザにかかっている確率はどれくらいか？？**

「インフルエンザにかかっているときに、陽性の確率は0.98だから、
かなり高い確率だと思うけれど？？」

例

・インフルエンザが流行していて、人口の**３％が罹患**している。

・検査薬は不完全で

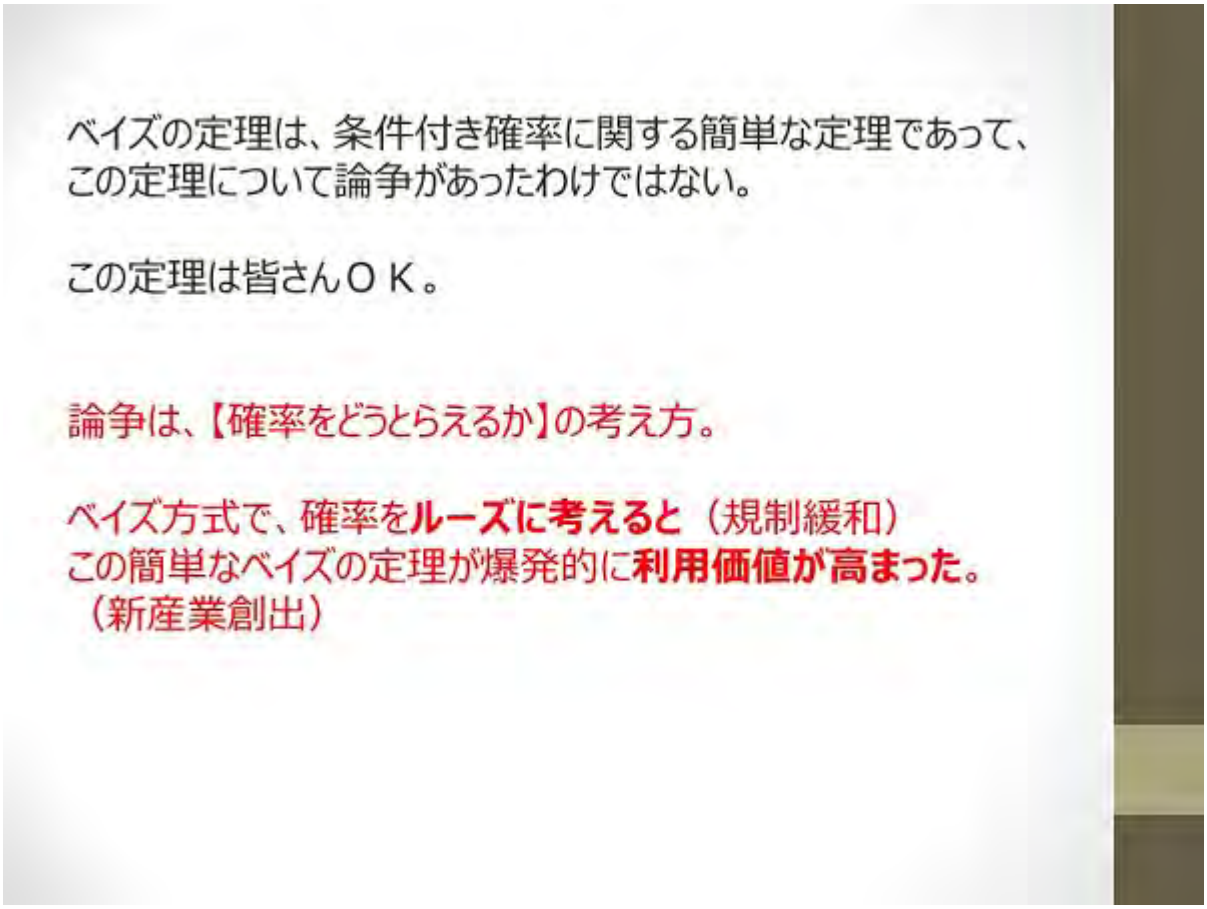
| | |
|---|---|
| インフルエンザにかかっていて陽性の確率 | 0.98 |
| インフルエンザにかかっていて陰性の確率 | 0.02 |
| インフルエンザにかかってなくて陽性の確率 | 0.05 |
| インフルエンザにかかってなくて陰性の確率 | 0.95 |

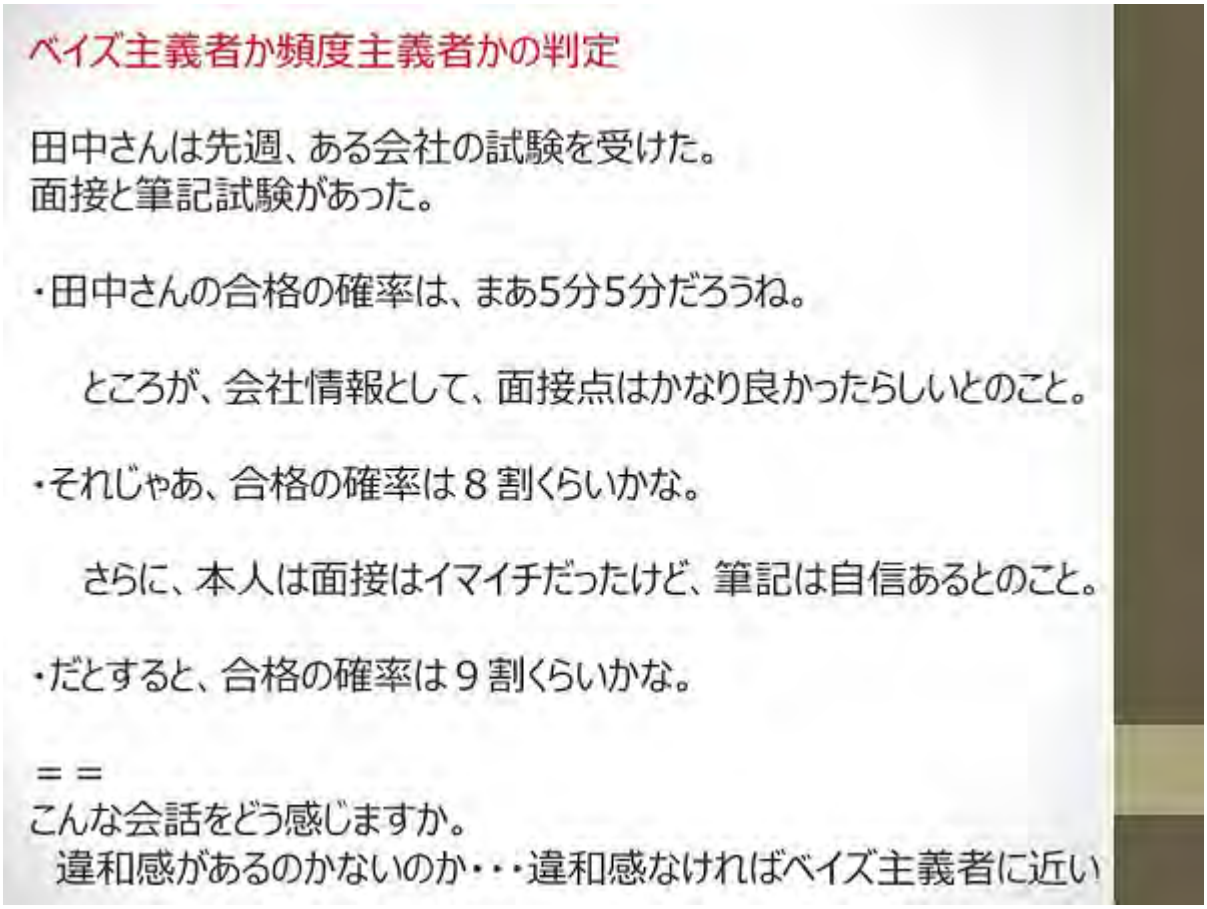
**さて、ある人が検査で陽性でした。**
**インフルエンザにかかっている確率はどれくらいか？？**

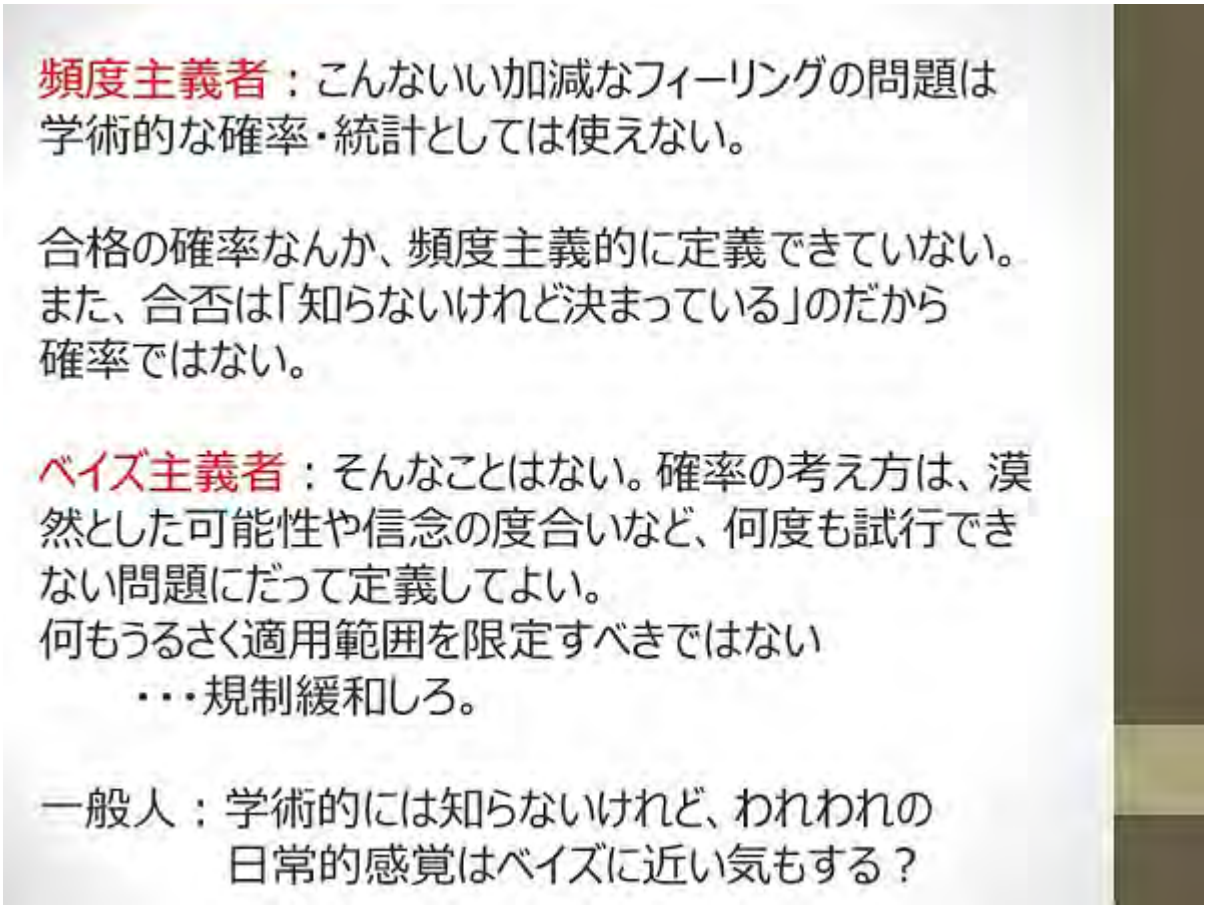
「インフルエンザにかかっているときに、陽性の確率は0.98だから、
かなり高い確率だと思うけれど？？」

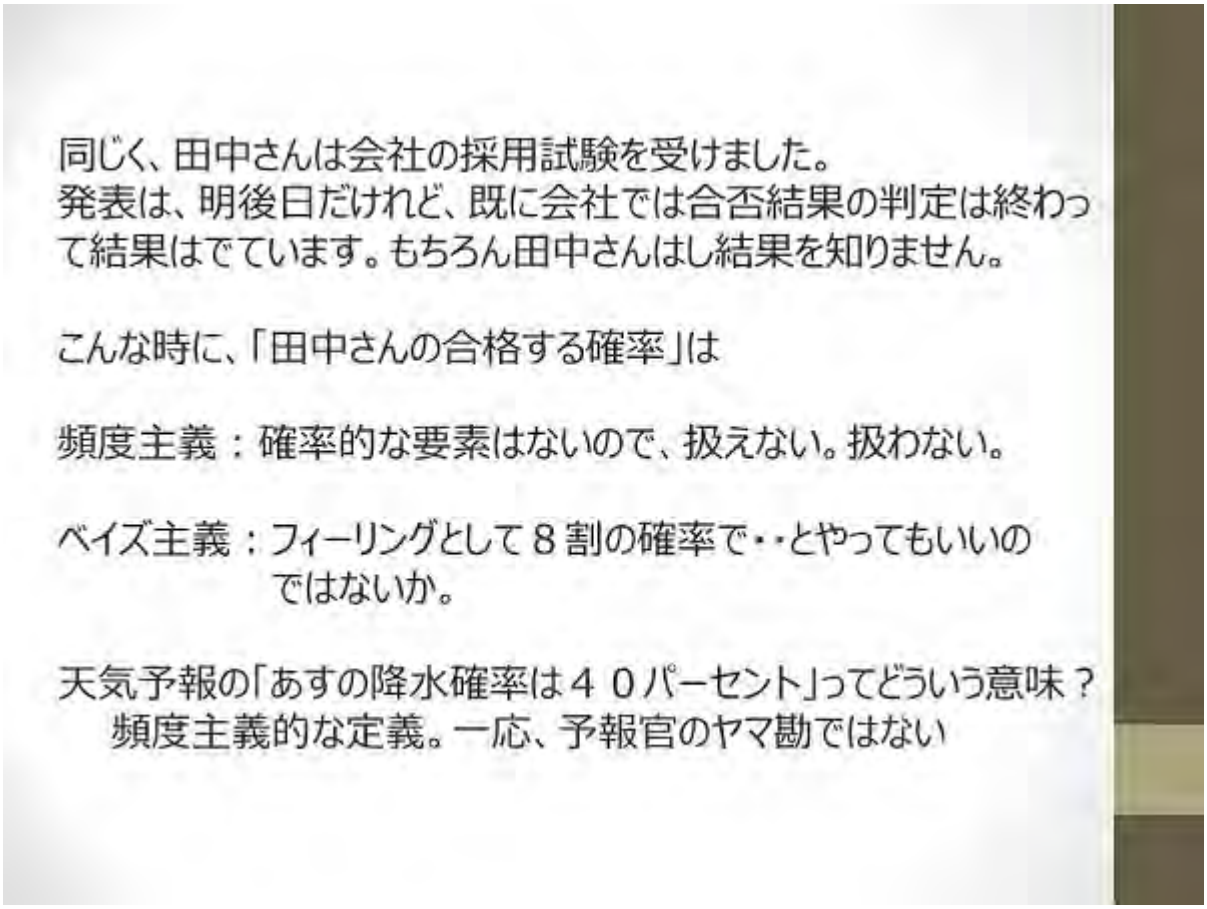
このくらい簡単なばあいは、ベイズの定理を使わないで、
図で簡単にやれる。意外に、インフルエンザの確率は小さい。

$0.03 \times 0.98/(0.03 \times 0.98 + 0.97 \times 0.05)=0.38$


ベイズの定理は、条件付き確率に関する簡単な定理であって、
この定理について論争があったわけではない。
この定理は皆さんＯＫ。
論争は、【確率をどうとらえるか】の考え方。
ベイズ方式で、確率をルーズに考えると（規制緩和）
この簡単なベイズの定理が爆発的に利用価値が高まった。
（新産業創出）


ベイズ主義者か頻度主義者かの判定
田中さんは先週、ある会社の試験を受けた。
面接と筆記試験があった。
・田中さんの合格の確率は、まあ5分5分だろうね。
ところが、会社情報として、面接点はかなり良かったらしいとのこと。
・それじゃあ、合格の確率は８割くらいかな。
さらに、本人は面接はイマイチだったけど、筆記は自信あるとのこと。
・だとすると、合格の確率は９割くらいかな。
＝＝
こんな会話をどう感じますか。
違和感があるのかないのか・・・違和感なければベイズ主義者に近い


頻度主義者：こんないい加減なフィーリングの問題は
学術的な確率・統計としては使えない。
合格の確率なんか、頻度主義的に定義できていない。
また、合否は「知らないけれど決まっている」のだから
確率ではない。
ベイズ主義者：そんなことはない。確率の考え方は、漠
然とした可能性や信念の度合いなど、何度も試行でき
ない問題にだって定義してよい。
何もうるさく適用範囲を限定すべきではない
・・・規制緩和しろ。
一般人：学術的には知らないけれど、われわれの
日常的感覚はベイズに近い気もする？


同じく、田中さんは会社の採用試験を受けました。
発表は、明後日だけれど、既に会社では合否結果の判定は終わっ
て結果はでています。もちろん田中さんはし結果を知りません。
こんな時に、「田中さんの合格する確率」は
頻度主義：確率的な要素はないので、扱えない。扱わない。
ベイズ主義：フィーリングとして８割の確率で・・とやってもいいの
ではないか。
天気予報の「あすの降水確率は４０パーセント」ってどういう意味？
頻度主義的な定義。一応、予報官のヤマ勘ではない

コイン投げの実験：コインの表が出る確率 θ

**頻度主義**

θは決まった値だが我々はそれを知らないので、実験で決めよう。

1000回投げて、表と裏の回数をカウントして、**θを推定しよう。**

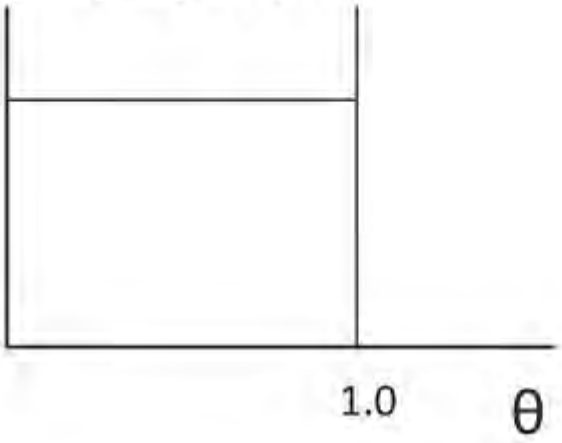
その推定値がどの程度正確なのかを見積もろう

コインの裏表が平等に出るかどうか（θ＝１／２かどうか）を
コイン投げの実験結果から検定しよう。

**ベイズ主義でやると**

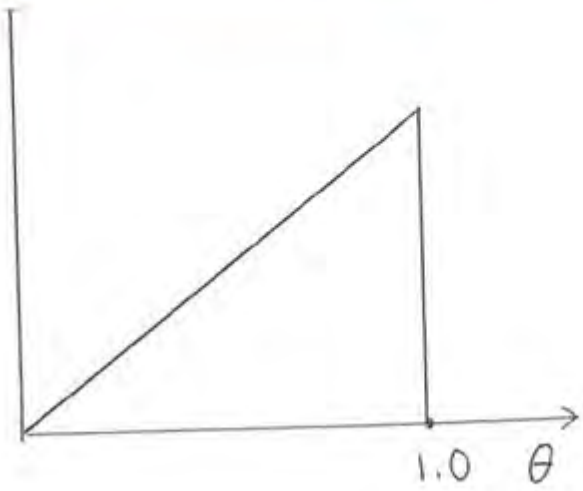
Θの値がわからないのだから、これを確率変数として扱い、確率分布を考えよう。
どんな確率分布かわからないので、まずは０～１の一様分布としておこう。

表の出る確率θ の分布



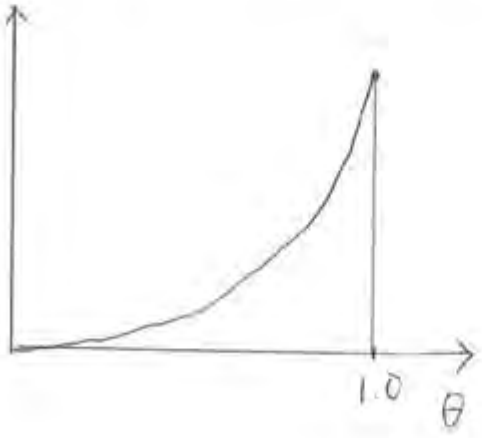
コインを１回目に投げたら「表」だった
「表」

確率分布をベイズの定理に従って修正



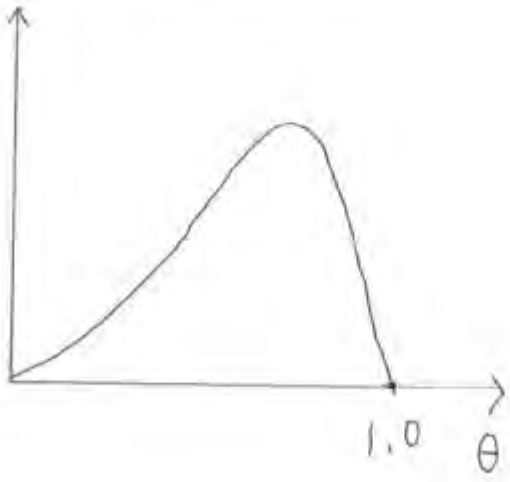

第二回目に投げたらまた「表」だった
「表表」

確率分布をベイズの定理に従って修正



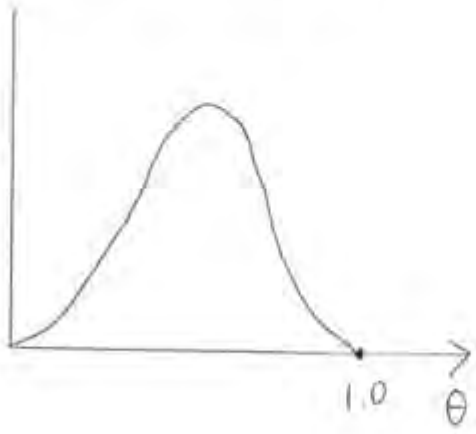
第三回目に投げたら「裏」だった
「表表裏」

確率分布をベイズの定理に従って修正



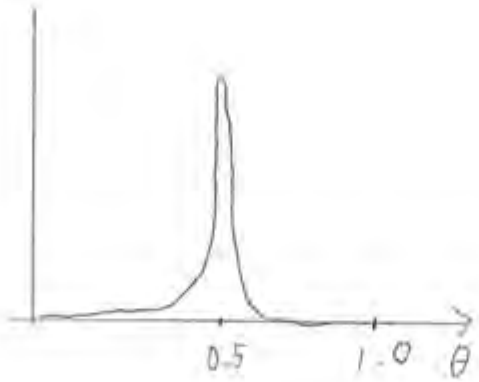

第四回目に投げたら「裏」だった
「表表裏裏」

確率分布をベイズの定理に従って修正



さらにコインを投げて「表」が１０２回、「裏」が９８回だった。

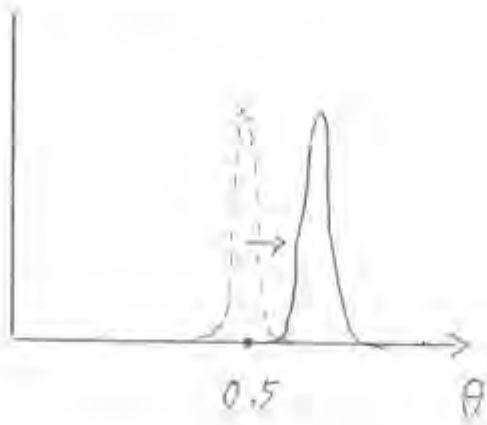
確率分布をベイズの定理に従って修正




何かの理由で、コインの表の出る確率が０．５から０．６に変わったとする。
コイン投げの実験を繰り返すと。

確率分布をベイズの定理に従って修正

自動的にピークが０．６のところに移動していく。
リアルタイム的な追随に有利！



例えば、日本国民の身長の平均値を調べたいとする。

H氏：日本人の平均身長は、何センチかわからないけれど決まっているのだから、**平均身長に確率の入る余地はない。**
日本人の中からランダムにサンプルをとって、身長を測りそのデータから平均身長を推定するのが統計の役割だ。

B氏：平均身長は決まっていても、わからないのだから、**確率変数として扱えばいいんだ。**日本人の平均身長はだいたいのところ、１６５センチくらいだろうから、１６５センチ付近にピークのある「平均身長の」確率分布を考えて、それがランダムサンプルの結果で徐々に精密に修正すればよい。

論点：
「日本人の平均身長」のように、**確定しているが知らないモノを母集団のパラメータ**として考えるか、**確率変数として考えるか→確率変数とするとベイズの定理が使える。**

頻度主義とベイズ主義の感覚の相違

H氏：
・キッチリした体系に載せたい。
・生物、農学などの現場で、比較的少数の標本データがとれて、それを詳細に分析するという古典的スタイルが基盤。

B氏：
・ソモソモ頻度主義的に扱えない、たった一度限りのことなども扱いたい・・・実務の現場・・・暗号、軍事、保険・・・。

・ネット社会になって、時々刻々新しいデータが流れ込む。
常に、リアルタイムで考え方をそれに合わせていかないといけない。
古典的スタイルでは間に合わない。
スパムメールやコンピュータウイルスは、常に新手が現れるので、常時それに対応する必要がある。

ネット社会では、ベイズ派の方が優勢になりつつある。
その場その場で、適当な方を選ぶということで折り合いをつけつつある。

古典統計学は、「農業試験場」や「遺伝学」などの、それほど多くない貴重なデータから何が言えるかを、慎重に取りだすためのマニュアルである・・・個別問題の形によってさまざまな匠の技のようなものがある。

現代では、情報過多で膨大なデータが流れていて、その中から何らかの判断を行わないといけない。ベイズ統計の方が、そのような状況にはフィットしている。

古典統計学にも優位な点は多い。

臨機応変に、その場その場の問題によって使い分ければいいのではないか・・・統計学者でもないわけだし。

実際の問題では、複雑な計算を要するが、いろいろな計算技術が開発されている（MCMC法など）。

いろいろな分野に浸透している。

ＥＮＤ


休憩中

以上