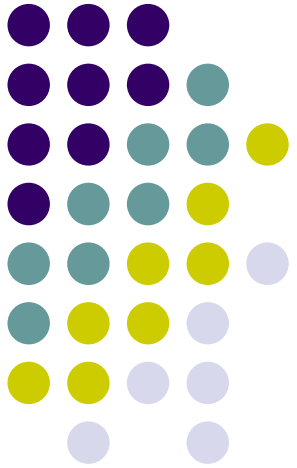
統計解析手法を使用したリスク分析へのアプローチ

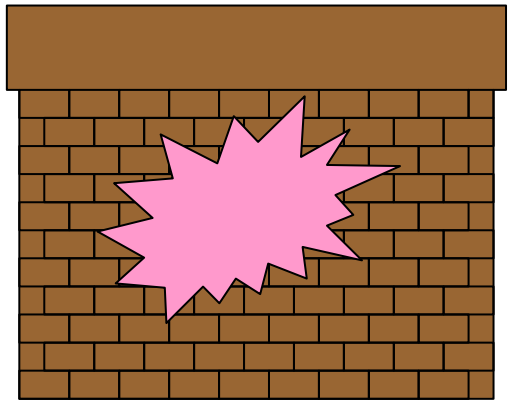
JNSA　リスク評価検討WG

2011/1/25 NSF2011 A3



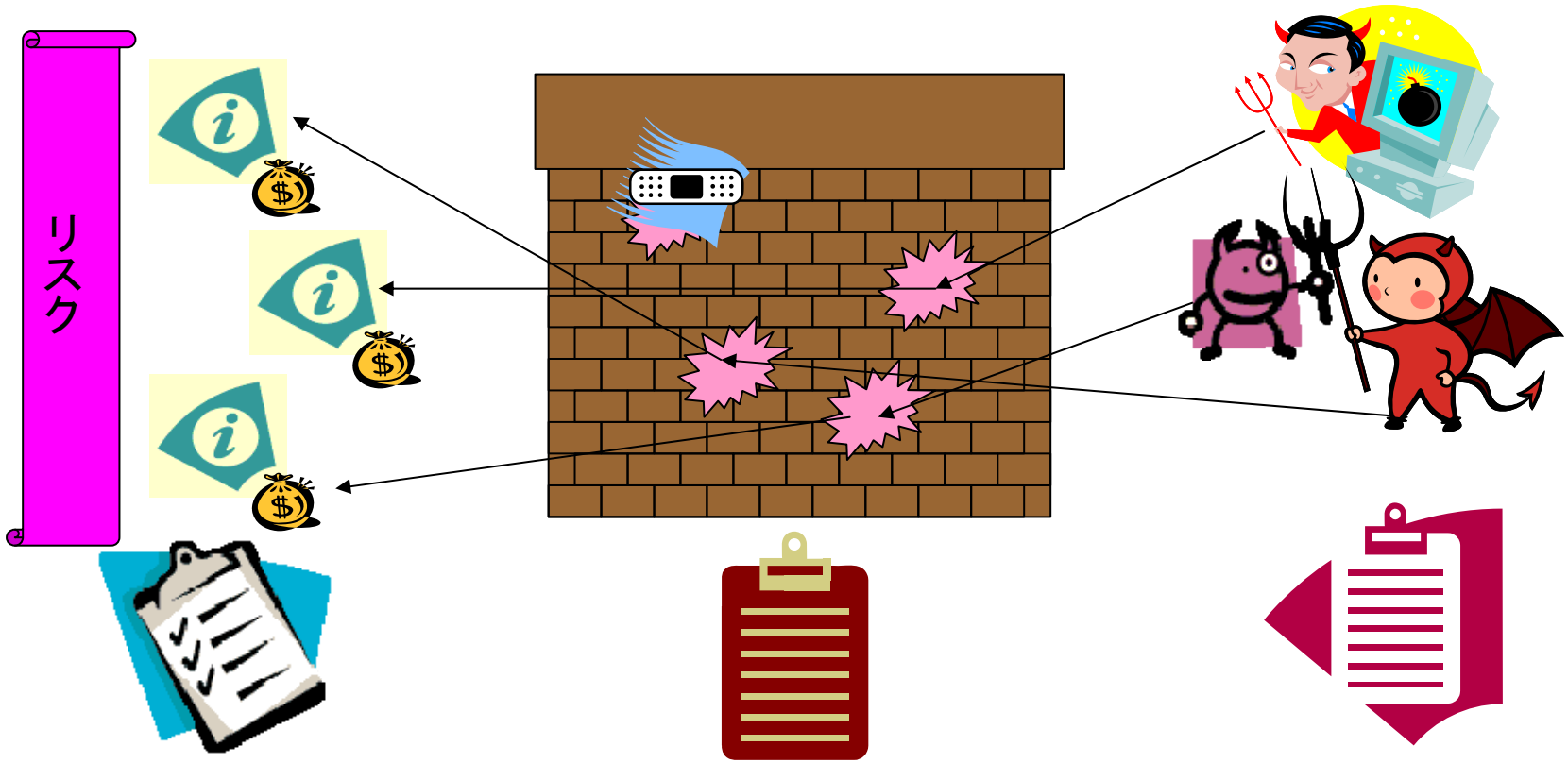
情報セキュリティにおける「リスク」

- 情報セキュリティにおける教科書的「リスク」



脅威　×　脆弱性　×　情報資産（価値）　＝　リスク

一般に行われているリスク分析



N段階評価

情報資産棚卸しと評価

N段階評価

脆弱性と対策の評価

N段階評価

脅威洗い出しと評価

従来型評価のメリット・デメリット

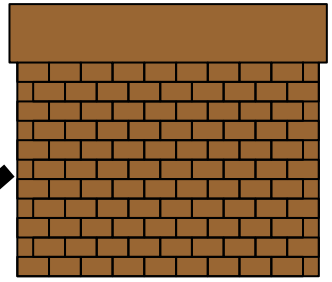
- ボトムアップ評価であるメリット
  - 個々の情報資産、脅威や脆弱性の所在と関連が明確にできる
  - ・・・つまり、効率よく（予防）対策を行うことができる
  - リスクの相対的な比較は可能であり、対策の優先付けも可能
- 準定性的（相対）評価であることのデメリット
  - リスクの絶対的な大きさを評価できない
  - ・・・つまり、（対策にもかかわらず）被害が出る可能性を予測できない
  - 予防や事後処理のために用意しておくべき費用を考える直接的な材料にはならないため、他のビジネスリスクと同じ土俵で評価ができない。（経営的には大きな不満点）

リスク定量化（計量）という難題

- リスク＝損失を発生する「可能性」
  - それはどれくらい大きなものなのか・・という問い
    - ビジネス上、最も重要なポイントである。
  - 「事故は発生するという前提」での予測が必要
    - これまでのセキュリティは、「予防」を主として考えられてきたが、実際は多くのインシデントが「発生」している。
    - インシデントの発生をゼロにすることは「不可能」という前提に立つ必要がある（発想の転換）

従来型評価における定量化

情報資産の価値
インシデントの影響評価



脅威の大きさや対策と
脆弱性の定量評価

インシデントの発生確率

対応コストの評価

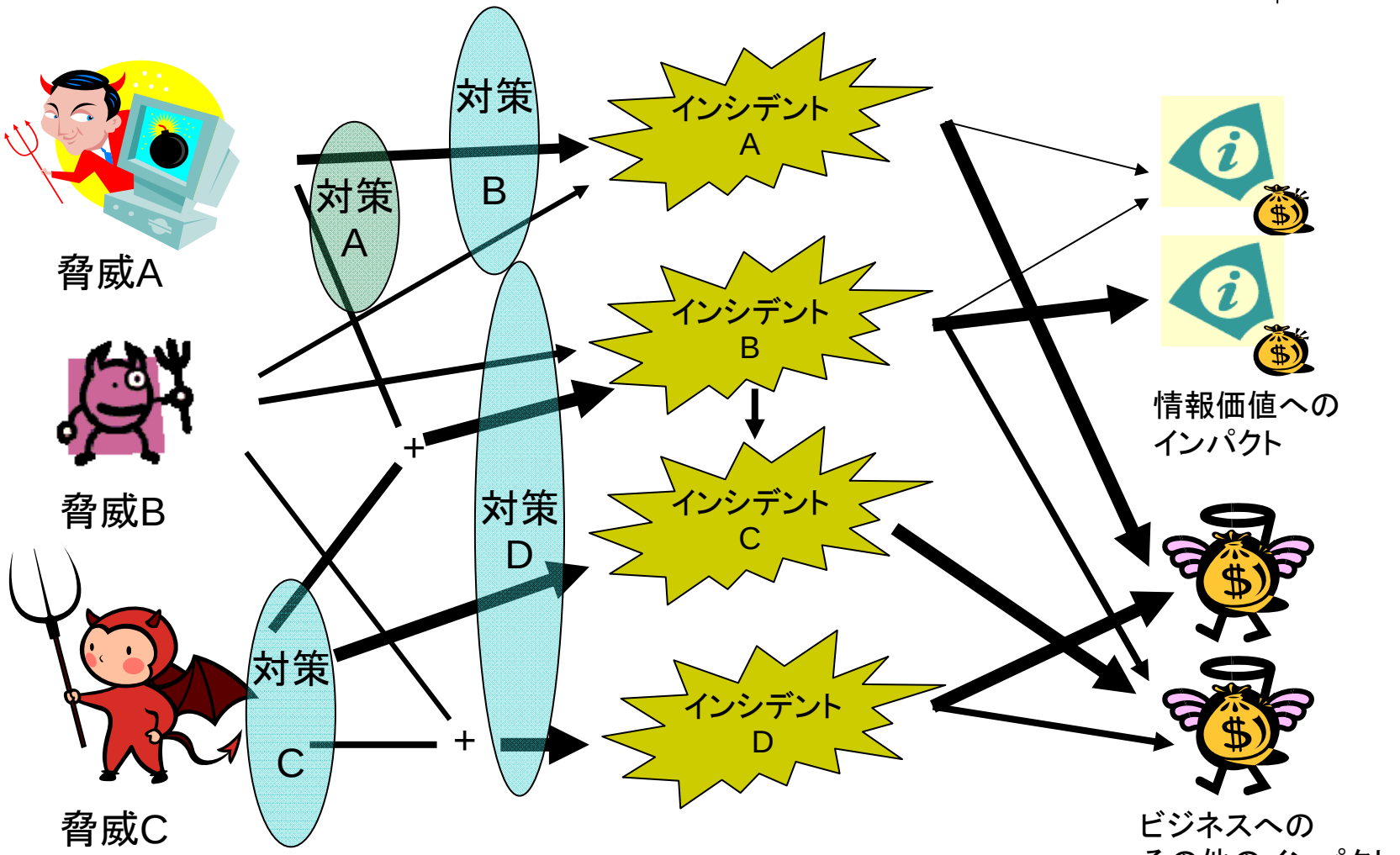
期待損失

理屈はわかるが、評価は
非常に難しい。

考えれば考えるほどモデル
は複雑化していく。

ボトムアップモデルの複雑化

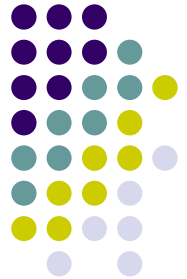
このすべてのパスについてリスクの総和を
計算できるのだろうか・・・・



山積みの課題

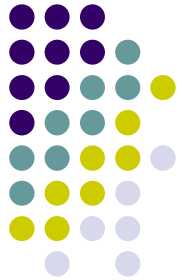
- 誤差の積み上げという問題
  - モデルを精緻にすればするほど、数値に精度が要求される
  - リスクの総量を求めると、その中に個々の数値の誤差がすべて積み上がってしまう
- リスク推定のためのデータが少ないという問題
  - 個々の脅威からたどってインシデント発生確率を求めるには、そのためのデータが少なすぎる
    - たとえば個々の組織では、得られないデータのほうが多い
    - 社会全体でも発生頻度が低いが、インパクトが大きなインシデントの評価は難しい

発想の転換



- 社長はリスクの総量を知りたい
  - 技術的な話や専門的な話は経営トップは興味なし
    - ・・と言い切ってしまうのは問題かもしれないが・・
  - 要するに、ビジネスリスクとしての期待損失をいくら見込んでおくべきなのかということ
  - それは、予防策に使うコストとどのように相関するのか、ということ
  - ただし、その根拠は「わかりやすく」説明する必要がある
    - ボトムアップ方式ではモデルが複雑すぎて説明も難しい
    - 多少大雑把でもいいから、より単純化されたモデルが必要
      - たとえば、それが一千万なのか一億なのか・・・というレベル（よりももう少しだけ細かく・・・）

つまり・・・・



- インシデント被害調査報告に人気がある理由
  - 実際に発生した事故の被害推定はわかりやすい
  - 傾向がよくわかるのでリファレンスとして使いやすい
- さらに・・・
  - 発生確率調査は、より直接的に使えそう
- でも、どのように使えばいいのか・・・
  - 報告で見えている傾向は本当に正しいのか？
  - 社会全体の傾向と自社の違いはどう評価しようか？
    - 同じ？、違う？、どれくらい違う？、何が違う？

リスク評価検討WGのスタンス

- 従来型のボトムアップ評価手法は必要
  - リスクの所在場所や原因、相対的な大きさを判断できるので、対策を考えるためには必須
  - ただ、定量化のためには、精緻なモデルと多くのデータの積み上げが必要で、まだ道のりは長い
- いっそ、総量を直接推定してみたら・・・
  - 一般のビジネスリスク評価は、どちらかといえばトップダウンで行われることが多い。
  - これまで被害調査で集めたデータや、今後の調査データをもとに、「自分の組織の」リスク総量を推定できないだろうか。

で、やっと本論・・・・・

- まずは、他のビジネスリスク評価で使われている方法や、こうした評価に使える統計解析技法を勉強してみよう（今年度の課題）
  - 企業における「オペレーショナルリスク」評価手法の勉強
  - さまざまな統計手法（確率分布と推定、検定などの方法論）の勉強とリスク評価への応用検討・試行
- 実際のデータへの応用（来年度課題）
  - インシデント発生確率調査、被害調査データから見える傾向の統計的検証
  - 自社の統計データを公的なデータで補完して精度を上げる、もしくは、公的なデータを自社のデータで補正する方法の検討

統計手法の勉強と試行

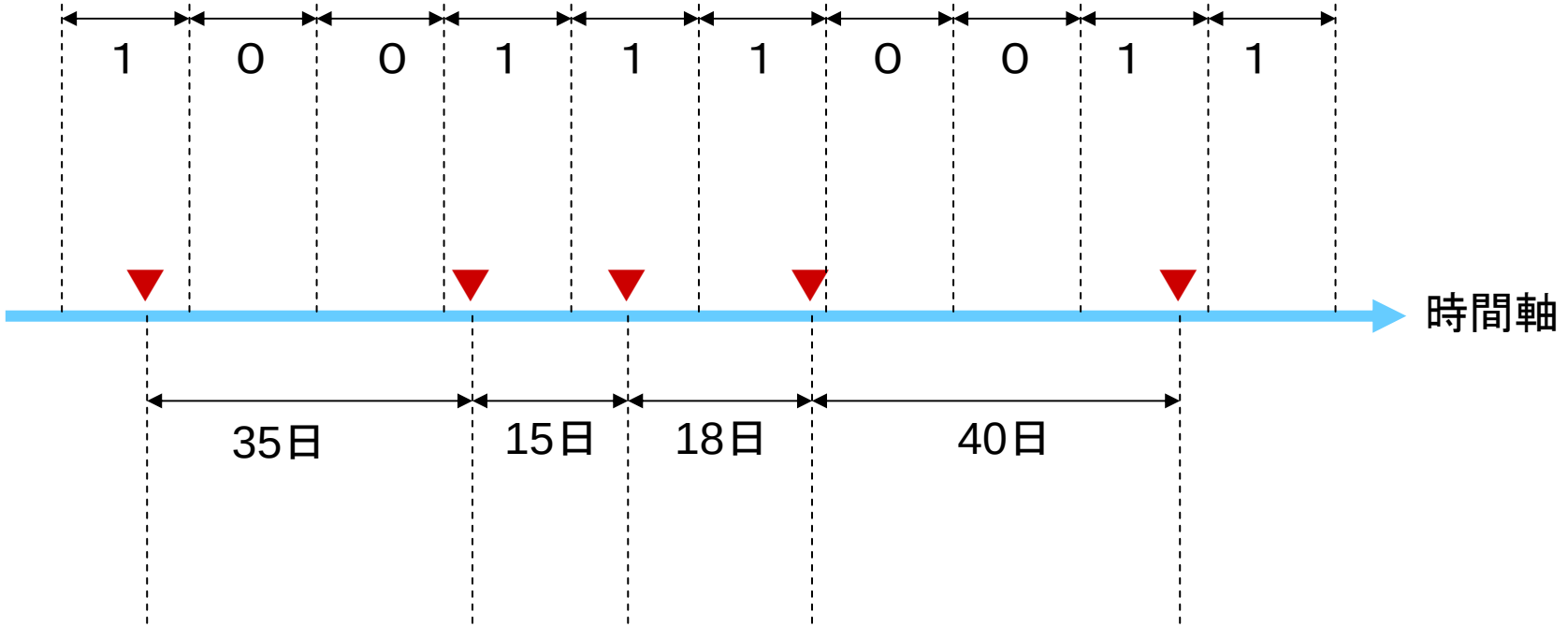
- インシデント被害調査WG過去データの解析
  - 統計手法の応用検討としての利用
  - データ利用方法を考えるための基礎作り
    - 被害調査WGとの間での協力関係の確立
- インシデント発生確率調査結果の解析
  - 調査結果に対する様々な統計処理の試行
  - 結果解釈の妥当性についての統計的検証の試行

被害調査データを使う前提

- 主にマスメディアや企業・組織の公表情報をもとにしていること
  - メディアによるバイアス、組織判断によるバイアスなどがかかっている可能性がある
  - インシデントとしては「氷山の一角」である可能性が高い
  - 個人情報漏洩のみを対象にしている
- それでも「使える」理由
  - メディア沙汰、不祥事の公表という事態が情報セキュリティインシデントおいて経営層が危惧する最大のポイントであること
  - 自社で、「公になるような（しなければいけないような）インシデント」が発生する可能性を考えるための材料として使えばいい
  - 情報「漏洩」のメカニズムには共通点も多いので、一般の情報漏えい事故にも拡大適用できるかもしれない（要検証）

インシデント発生間隔分析

発生頻度　　一定の期間に発生する回数



発生間隔　　インシデントから次のインシデントまでの間隔

なぜ発生間隔に着目したか

- （理由１）そう頻繁に発生する事象ではない
  - 頻度を見る場合、期間を短く取れば0が多くなる。長く取れば粒度が荒くなる
  - （ポアソン分布で近似することもできるのだが）
- （理由２）いつかは必ず発生する（仮説）
  - インシデントをゼロにすることはできない
  - 学習・忘却曲線的な周期または減衰曲線的なパターンがあるかもしれない
    - しかも、個々の組織だけではなく、社会的にも
    - 「喉元過ぎれば・・・・」のたとえ
  - つまり、地震や火山噴火に特性が類似している
    - これらの災害では、発生間隔に着目している
    - 長期間発生していないほど発生が近づいている・・・という見方

被害調査データと発生間隔

- 業種ごとに、インシデントの発生間隔を見てみた
  - 業種によって偏りが出るだろうという「仮説」
- 平均発生間隔を区間推定してみた
  - ある時期のデータの平均発生間隔、分散値（サンプル平均、分散）から本来の平均間隔の推定範囲を算出
- 業種ごとの偏りの検証
  - 分散分析と多重比較分析による業種ごとの有意差検定

業種全体で「発生間隔」を考える意味については、まだ議論が尽くされていないが、まずは統計手法の応用検討という意味で実施した。

2008年～2009年度のデータから

業種	その他 サービス業	金融 保険業	小売業	情報 通信業	教育・学習 支援業	電気・ガス 水道業*	製造業*	医療 介護	不動産業*
N	87	158	72	94	177	67	59	90	71
平均間隔	3.92	2.24	4.76	3.74	1.94	10.27	12.02	3.82	9.83
不偏分散	17.1396	10.5625	23.3289	15.9201	4.2849	112.9969	171.8721	12.25	147.8656
標準偏差	4.14	3.25	4.83	3.99	2.07	10.63	13.11	3.5	12.16
95%上限	4.79	2.75	5.88	4.55	2.55	12.79	15.36	4.55	12.66
95%下限	3.05	1.73	3.65	2.94	1.64	7.75	8.67	3.10	7.00

たとえば、金融・保険業全体では、平均2．24日に1回という平均値になったが、サンプル数87での分散値から不偏分散を計算し、母平均の95%信頼区間を求めると、3.05日～4.79日となった。（平均±1.96σ/√N）（但し、これは正規分布を仮定したものなので、数学的な意味では再検討が必要かもしれない）

ここで、この母平均の推定区間と、標準偏差（不偏分散値をもとにしたもの）の関係を見てみると、標準偏差の値が平均値の95%推定区間内に含まれており、この二つの値が等しいという仮説を危険率5%で棄却できないことがわかる。一般に、こうした事故の発生間隔は指数分布で近似できることが多いが、指数分布の特性として、係数λ（$f(x) = \lambda e^{-\lambda x}$）に対して、平均1／λ、分散1／$\lambda^2$ の関係が知られており、これを裏付ける結果とも考えられる。

但し、これは業種全体での値であり、これを個々の企業でどう解釈するかについては今後の検討画題である。

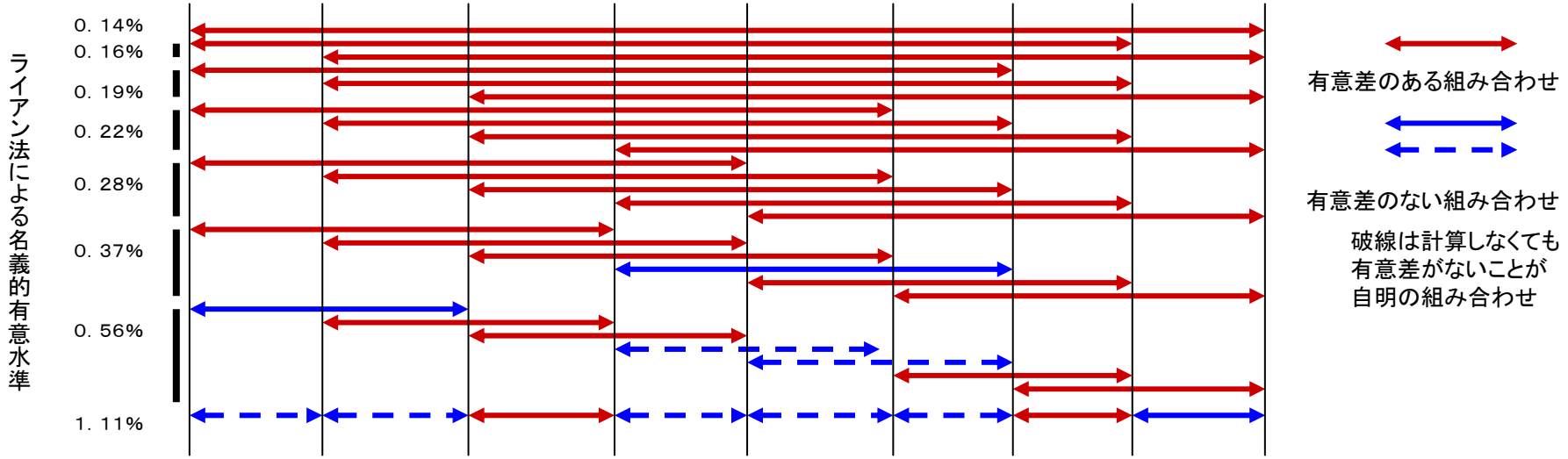
業種間の差異分析

- 一見して差異があることは明らかに見えるが
  - 分散分析（業種別の偏り（効果）と誤差的なばらつきの比較による検定の一種）でも、業種間差異は明らかに見える
  - では、各業種間の差異はどうだろうか
    - 多重比較分析が必要
      - 業種間の有意差を個別に検定
      - ライアンの方法による検定水準の調整を行って、全体として5%危険率を保てるようにした

多重比較結果

先の結果を平均値の大きい順に並べ替え

ランク	1	2	3	4	5	6	7	8	9
業種	製造業*	電気・ガス水道業*	不動産業*	小売業	その他サービス業	医療介護	情報通信業	金融保険業	教育・学習支援業
N	59	67	71	72	87	90	94	158	177
平均間隔	12.02	10.27	9.83	4.76	3.92	3.82	3.74	2.24	1.94
標準偏差	13.11	10.63	12.16	4.83	4.14	3.5	3.99	3.25	2.07
分散	171.87	112.99	147.86	23.32	17.13	12.25	15.92	10.56	4.28



インシデントの発生確率

- 発生間隔の分布が指数分布に従うとして
  確率分布関数 $f(x) = \lambda e^{-\lambda x}$

  但し：　平均値E(X)＝1／λ　　分散V[X]＝1／λ^2

  平均値信頼区間の下限と上限について、それぞれλ_1 λ_2を求めてみる

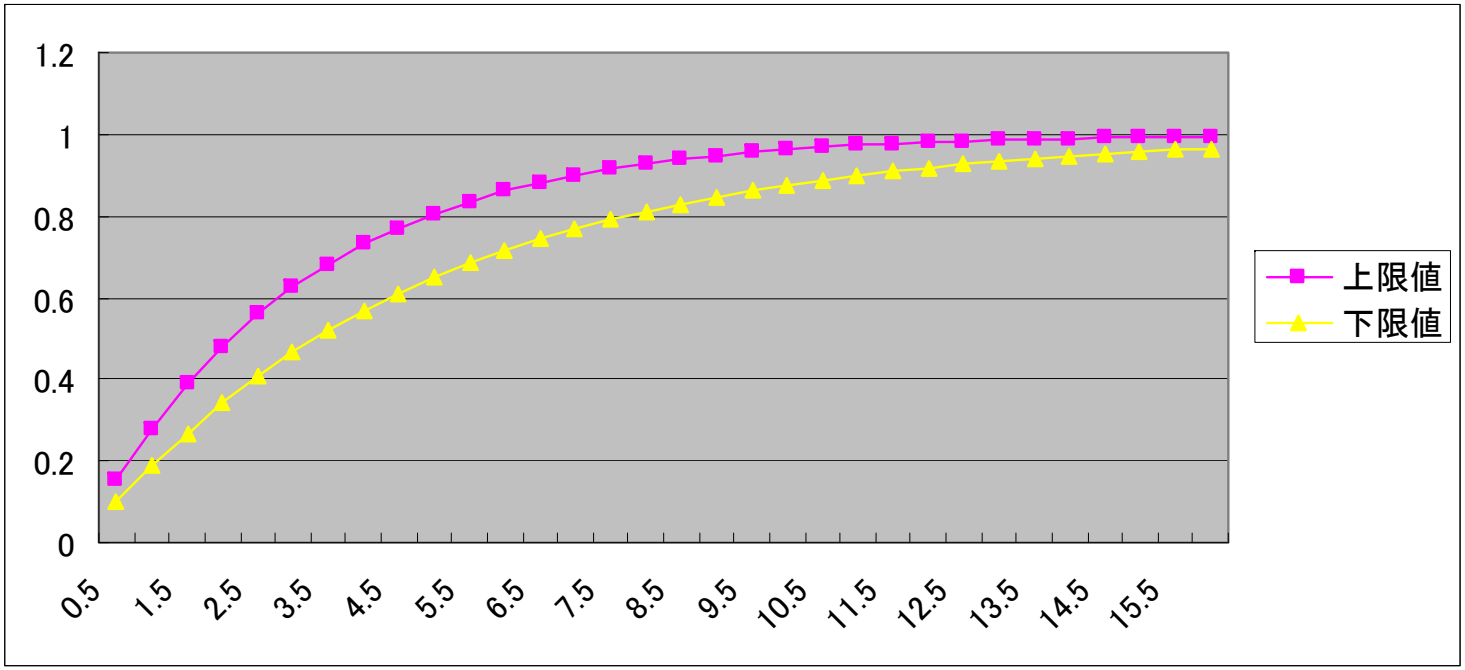
  λ＝1／E(X)　　この場合の分散はV[X]＝1／(E[X])^2

  あるインシデントが発生してからx日で次のインシデントが発生する確率

$$P(x)= \int_0^x \lambda e^{-\lambda x}\ dx = 1 - e^{-\lambda x}$$

  について、2つのλ_1，λ_2を使って計算してみると、発生確率の区間推定のようなことができるはず。

その他サービス業の場合



平均値区間 3.05 ≦E(X)≦4.79　　λ_1=0.151　λ_2=0.099

たとえば、この場合、あるインシデントが発生してから10日間以内に次のインシデントが発生する確率は、おおむね90%から97%程度の範囲にあると言える。

これを個々の組織でどう使うか

- 業種全体での発生間隔と個々の組織の発生間隔を個々の組織の発生間隔に落とす方法はないだろうか
  - 数百万人レベルでは平均化されてしまうような、企業や個人によるばらつきが、無視できなくなる可能性がある。
- 業種データをもとに、それになんらかの補正をかけるか、もしくは、個々の組織が独自でデータを集め、少ないデータで精度を上げる方法を考えるか、どちらがよいかの結論はまだ出ていない。
  - 最終的には個々の組織でのデータ蓄積は必須だが、それには長年の継続的データ収集を必要とする。精度はさほど高くなくても、数年程度のデータ蓄積にかわる基準データを提供できれば、それを使いながら、徐々に自社データに置き換えていくことが可能になるのではないか。（WGレビューでの議論から）

ためしにひとつの試算

- その他サービス業における考察
  - 就労者数：463万人（平成21年度：政府統計）
  - 従業員数1000人の企業
    - 係数A＝4630000/1000＝4630
  - この業種における平均発生間隔は3．92日なので、先の仮定が成り立つとして、この企業における（新聞沙汰になるような）インシデントの平均発生間隔は
    - 4630×3．92＝18149．6日＝約49年
  - これが10000人規模の企業だと、約5年に1度となる。

＊この方法では、発生間隔が直線的に従業員数に比例してしまうため、小規模な企業や特に大きな企業では実態とかけはなれた値になってしまう可能性が高い。
＊やはり、それほど単純にはいかない・・・・という例。

発生間隔分析のまとめ

- 指数分布が前提であれば、決めるべき係数は「平均値」のみ。
  - 確率計算が単純化できる上、この関数は計算が容易。
  - しかし、それゆえに平均値の精度が問題になる。
    - 少ないデータで精度よく平均値をはじきだす方法はあるだろうか（なんとなく難しそうな気もする・・・）
    - 業種別平均を使って、まずは、大きく外れない（この定義も難しいが）概算値を出せないか（それを初期値として、その後のデータを積み重ねていくことで、精度向上が望める）

＊このあたりが、これからのWGにおける課題のひとつ。

インシデント発生確率調査の検証

- アンケート調査方法の検証
  - サンプル数から母集団の統計値との関係を検証
    - この結果が、どの程度、日本全体にあてはまるだろうかということの統計的な検証
- 個人の特性に対する依存性に着目
  - 個人によって、どの程度インシデントの可能性はばらつくのか
  - そのばらつきの原因は何か
    - 相関関係の検証など

母比率推定

アンケートで見られた差異は本当か？

Q1 これまでに会社貸与の携帯電話、私物の携帯電話を社内・社外を問わず、紛失した・盗難にあったことがありますか。（お答えはいくつでも）

	回答数	%
全　　体（N）	4,884	100.0
業務データが入った会社貸与の携帯電話を紛失した・盗難にあったことがある	184	3.8
業務データが入った私物の携帯電話を紛失した・盗難にあったことがある	204	4.2
業務データが入った会社貸与の携帯電話を紛失しそうになったことがある	269	5.5
業務データが入った私物の携帯電話を紛失しそうになったことがある	288	5.9
業務データが入っていない会社貸与の携帯電話を紛失した・盗難にあったことがある	80	1.6
業務データが入っていない私物の携帯電話を紛失した・盗難にあったことがある	305	6.2
会社貸与や私物の携帯電話を紛失した・盗難にあったことがない	3,906	80.0

$\sigma = \sqrt{p(1-p)/n}$

			95%信頼区間		信頼度95%推定値	
p	σ	1.96σ	上限値	下限値	上限値	下限値
0.038	0.002736	0.005362	0.043362	0.032638	4.3%	3.3%
0.042	0.00287	0.005626	0.047626	0.036374	4.8%	3.6%
0.055	0.003262	0.006394	0.061394	0.048606	6.1%	4.9%
0.059	0.003372	0.006608	0.065608	0.052392	6.6%	5.2%
0.016	0.001795	0.003519	0.019519	0.012481	2.0%	1.2%
0.062	0.003451	0.006763	0.068763	0.055237	6.9%	5.5%
0.8	0.005724	0.011218	0.811218	0.788782	81.1%	78.9%

たとえば、質問1の回答率3．8%が、実は4．2%であり、質問2の回答率と同じである可能性が2．5%以上存在する。逆もまたしかり

実際はこれだけの幅があってもおかしくない

「PC紛失は酒場より社内」は本当？

検定するまでもなく一目瞭然でした

PC紛失　回答総数＝100

母比率推定
95%

	回答数	比率p	1−p	x=p(1−p)/n	1.96√x	95%下限値	95%上限値
勤務中、社内で無くした	29	0.29	0.71	0.002059	0.088937	0.20	0.38
通勤中、勤務で移動中に、タクシー、電車、飛行機などの乗り物に置き忘れた	24	0.24	0.76	0.001824	0.083708	0.16	0.32
取引先、出張の宿泊先、セミナ会場など、勤務中に滞在した施設で無くした	14	0.14	0.86	0.001204	0.068009	0.07	0.21
自宅、プライベートの出先で無くした。	8	0.08	0.92	0.000736	0.053173	0.03	0.13
飲酒して酔っているときに、飲食店、タクシー、電車などの交通機関で無くした	4	0.04	0.96	0.000384	0.038408	0.00	0.08
その他の紛失	0	0	1	0	0	0.00	0.00
通勤中、勤務中にひったくり、置き引き、車上荒らしなどの盗難にあった	6	0.06	0.94	0.000564	0.046547	0.01	0.11
自宅、プライベートの出先で盗難にあった	9	0.09	0.91	0.000819	0.056092	0.03	0.15
その他の盗難	2	0.02	0.98	0.000196	0.02744	0.00	0.05
いつ、どこで無くなったのか分からない	4	0.04	0.96	0.000384	0.038408	0.00	0.08

一方で携帯電話は？

携帯電話紛失　5%両側検定で、比率に差がないという仮説を棄却できません
PCを宴席に持ち込まないことは常識化しているが携帯は持ち歩かざるを得ないから？　母比率推定95%

	回答数	比率	1−p	x=p(1−p)/n	1.96√x	95%下限値	95%上限値
勤務中、社内で無くした	17	0.17	0.83	0.001411	0.073624	0.10	0.24
通勤中、勤務で移動中に、タクシー、電車、飛行機などの乗り物に置き忘れた	29	0.29	0.71	0.002059	0.088937	0.20	0.38
取引先、出張の宿泊先、セミナ会場など、勤務中に滞在した施設で無くした	11	0.11	0.89	0.000979	0.061326	0.05	0.17
自宅、プライベートの出先で無くした。	17	0.17	0.83	0.001411	0.073624	0.10	0.24
飲酒して酔っているときに、飲食店、タクシー、電車などの交通機関で無くした	11	0.11	0.89	0.000979	0.061326	0.05	0.17
その他の紛失	0	0	1	0	0	0.00	0.00
通勤中、勤務中にひったくり、置き引き、車上荒らしなどの盗難にあった	2	0.02	0.98	0.000196	0.02744	−0.01	0.05
自宅、プライベートの出先で盗難にあった	5	0.05	0.95	0.000475	0.042717	0.01	0.09
その他の盗難	0	0	1	0	0	0.00	0.00
いつ、どこで無くなったのか分からない	8	0.08	0.92	0.000736	0.053173	0.03	0.13

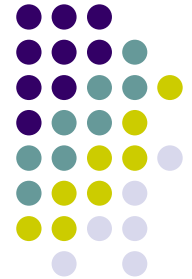
「リスクの高い人」は存在するのか？

サンプル数の少ない「紛失経験者」側の比率の母比率区間推定（95%）をして全体と比較してみた。
やはり、全体に比べて、紛失経験者は誤送信の可能性も高いという仮説は否定できない。

		全体(N)	メールを誤った宛先へ送信したことがある		宛先ではならない他人のメールアドレスが見えるように送信したことがある		機密情報など誤って記入したり、添付したりして送信したことがある		FAXで誤った宛先へ送信したことがある		FAXで間違った文書を送信したことがある		メールを誤送信したことはない		FAXを誤送信したことはない	
1	業務データが入った会社貸与の携帯電話を紛失した・盗難にあったことがある	184	129	70.1%	65	35.3%	42	22.8%	134	72.8%	69	37.5%	40	21.7%	34	18.5%
2	業務データが入った会社貸与の携帯電話を紛失した・盗難にあったことがない	4,700	2,916	62.0%	405	8.6%	189	4.0%	2,523	53.7%	978	20.8%	1,506	32.0%	1,570	33.4%
3	業務データが入った会社貸与のパソコンを紛失した・盗難にあったことがある	148	115	77.7%	61	41.2%	39	26.4%	110	74.3%	62	41.9%	20	13.5%	25	16.9%
4	業務データが入った会社貸与のパソコンを紛失した・盗難にあったことがない	4,736	2,930	61.9%	409	8.6%	192	4.1%	2,547	53.8%	985	20.8%	1,526	32.2%	1,579	33.3%
5	業務データが入った会社貸与のUSBメモリを紛失した・盗難にあったことがある	146	122	83.6%	66	45.2%	43	29.5%	120	82.2%	68	46.6%	13	8.9%	14	9.6%
6	業務データが入った会社貸与のUSBメモリを紛失した・盗難にあったことがない	4,738	2,923	61.7%	404	8.5%	188	4.0%	2,537	53.5%	979	20.7%	1,533	32.4%	1,590	33.6%

N	n1		n2		n3		n4		n5		n6		n7	
184	129	0.701	65	0.353	42	0.228	134	0.728	69	0.375	40	0.217	34	0.185
		0.767		0.422		0.289		0.793		0.445		0.277		0.241
		0.635		0.284		0.168		0.664		0.305		0.158		0.129
148	115	0.777	61	0.412	39	0.264	110	0.743	62	0.419	20	0.135	25	0.169
		0.844		0.491		0.334		0.814		0.498		0.190		0.229
		0.710		0.333		0.193		0.673		0.339		0.080		0.109
146	122	0.836	66	0.452	43	0.295	120	0.822	68	0.466	13	0.089	14	0.096
		0.896		0.533		0.368		0.884		0.547		0.135		0.144
		0.775		0.371		0.221		0.760		0.385		0.043		0.048

今後の方向性（遠大な計画？）



- 各種の公的データから業種別の傾向（業種別確率分布）を抽出（11年度）
  - 情報漏えい関連だけではなく、マルウエア感染など他のデータについても検討
- 発生確率調査をもとに、公的データを個人や環境的な属性で補正する方法を検討（できれば11年度）
- 被害額と発生確率との相関関係は？
  - 意外なことに、被害調査データから明らかな相関は見えない。（独立したパラメータとして扱える可能性を検証したい）
- 最終的に、JNSAから提供する各種統計値を自社データで補正する方法まで検討したい

参加をお待ちしています

- 求む、分析の「アイデア」
- 求む、「データ」
- 求む、「算数使い」
- 求む、「実験台」（笑）
- 求む、「突っ込み」（笑）
- とにかく、求む「参加者」・・・・・

ご清聴ありがとうございました。