【データの偏りを見抜く】

見えていない情報に気づく

データ分析にとりかかる前に，ちょっと考えてみましょう。もし，これから分析するために集めた「データ」そのものに「偏り」があったとしたら…分析をしてもうまくいきません。このような集めたデータの偏りは（　　）と呼ばれ，いろいろなものが存在します。これらは隠れてしまいやすいもので，見抜く力が必要です。

例１．アメリカの戦闘機を強くした統計学

第２次世界大戦では，飛行機がとても活躍しましたが，戦闘機は撃墜されれば機体が助かることはまずありません。かと言って装甲を厚くすれば重くなって速度が遅くなって不利になってしまいます。そこで，厚くする場所と薄くする場所を選ばなければならないわけです。そのために，帰還した戦闘機の弾痕調査を実施して分析しました。

［**考えてみよう１**］図１のような弾痕パターンが見られたとき，他より装甲を厚くすべき場所を斜線で図２に示してみましょう。

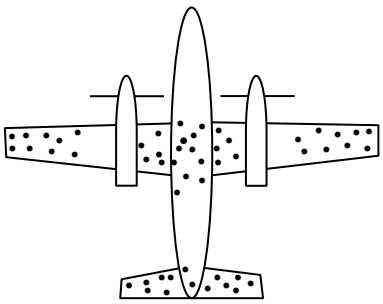

図１

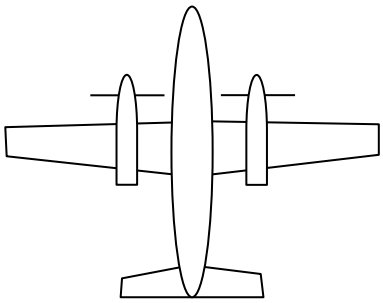

図２

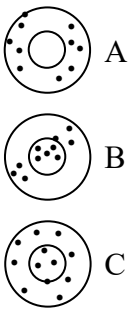

図３

考察①　動く丸い標的に銃を乱射したときの弾痕の予想は，　　のはず
→考察②　飛行機も標的と同じであるとすると，　　被弾するはず
→考察③　弾痕の　　場所に被弾したら，帰還できなかったのではないか
→結論　　装甲を厚くすべき場所は，弾痕の　　部分である。

●（　　）だけのデータを集めたため，（　　）のデータが全く集まっていない状態になっている。このような偏りを（　　）という。

［**回答課題５**］次のデータ分析結果について，偏りを見抜いて説明しよう。（Classroom 質問）
「動物病院への調査データによると，ネコは５階以下の高さから落ちた場合よりも，
６階以上の高さから落ちた場合の方が回復する割合が高い。終端速度に達して落下姿勢が安定したと考えられる。」

例２．ドイツを愛するハズレないタコ

2010 年に一匹のタコが有名になったそうです。FIFA ワールドカップ 2010 南アフリカ大会で，ドイツ代表の試合８つの勝敗をすべて的中させたというのです。餌と一緒に置いた２つの国旗のどちらへ行くかで決めたそうです。

［**考えてみよう２**］８つの試合の勝敗をすべて的中させる確率はどれくらいなのか，計算して分数で求めてみましょう。（引き分けは考えません。）

●これがとても珍しいことと言ってよいのか…というのは前提条件によるところがある。ワールドカップの勝敗は世界中の多くの人が関心を寄せ，同じように動物に勝敗を占わせていたのが（　２５６　）匹以上いたとすると，少なくとも 1 匹以上は確率的に的中するはずである。このとき，多くの（　　）**存在がある**，ならば確率的に珍しいとは言えなくなる。これも一つの**バイアス**と言える。

例３．食い違った２つの大統領選挙予想

1936 年のアメリカ大統領選挙においては，民主党ルーズベルト大統領と共和党ランドン候補が争っていました。このとき，２つの世論調査の結果も対立しました。

①「リテラリー・ダイジェスト誌」…過去 5 回の大統領選挙結果を当てており，サンプル数は 200 万人。ランドン候補が 57%の支持で有利と発表。

②「ギャラップ（当時はアメリカ世論研究所)」…当時は新興の調査会社であり，サンプル数は 3000 人。ルーズベルト大統領が 54%の支持で有利と発表。

結果はルーズベルト大統領の圧勝となり，世間を驚かせました。

［**考えてみよう３**］当時は①の調査結果が支持されていました。そのように信頼性が高いと思われていた理由を 2 点，簡潔に「～ので，～と思われていたから」という文で説明してみましょう。

・　　と思われていたから

・　　と思われていたから

●サンプル選びの違い…①と②ではサンプルの選び方に違いがあった。

①自社の雑誌読者，電話や自動車を所有する人を対象に得たサンプル

②都市の男女，農村の男女，中間層の男女…等の**層に細分化**し有権者数に合わせたサンプル

①のサンプルは結果として富裕層を中心に選んだことになり，それまでは母集団とあまり違わなかったものが，経済状況等の社会情勢の変化によって当てはまらなくなったと考えられる。②のようなサンプルの集め方は（　　）という。

●サンプル選びの方法…人が選ぶと頼みやすさ等で偏りがでることから，コンピュータを用いた（　　）**法**もしくは（　　）**法**と呼ばれる方法が利用されている。その一つが無作為に電話番号を生成して電話して調査を依頼する**RDD（Random Digit Dialing）法**であるが，これも間違いない方法とは言い切れない。

【データの偏りを見抜く】

見えていない情報に気づく

データ分析にとりかかる前に，ちょっと考えてみましょう。もし，これから分析するために集めた「データ」そのものに「偏り」があったとしたら…分析をしてもうまくいきません。このような集めたデータの偏りは（　**バイアス**　）と呼ばれ，いろいろなものが存在します。これらは隠れてしまいやすいもので，見抜く力が必要です。

例１．アメリカの戦闘機を強くした統計学

第２次世界大戦では，飛行機がとても活躍しましたが，戦闘機は撃墜されれば機体が助かることはまずありません。かと言って装甲を厚くすれば重くなって速度が遅くなって不利になってしまいます。そこで，厚くする場所と薄くする場所を選ばなければならないわけです。そのために，帰還した戦闘機の弾痕調査を実施して分析しました。

［**考えてみよう１**］図１のような弾痕パターンが見られたとき，他より装甲を厚くすべき場所を斜線で図２に示してみましょう。

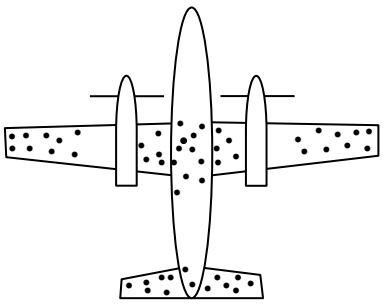

図１

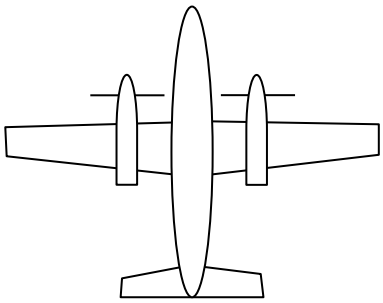

図２

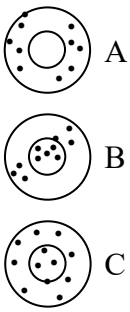

図３

考察①　動く丸い標的に銃を乱射したときの弾痕の予想は，　*図３-C*　のはず
→考察②　飛行機も標的と同じであるとすると，　*まんべんなく*　被弾するはず
→考察③　弾痕の　*ない*　場所に被弾したら，帰還できなかったのではないか
→結論　装甲を厚くすべき場所は，弾痕の　*あまりない*　部分である。

●（　*生存者*　）だけのデータを集めたため，（　*死者*　）のデータが全く集まっていない状態になっている。このような偏りを（　**生存者バイアス**　）という。

［**回答課題５**］次のデータ分析結果について，偏りを見抜いて説明しよう。(Classroom 質問)
「動物病院への調査データによると，ネコは５階以下の高さから落ちた場合よりも，
６階以上の高さから落ちた場合の方が回復する割合が高い。終端速度に達して落下姿勢が安定したと考えられる。」

６階以上の高さから落ちたネコには，病院に運ばれることなく亡くなったものが多くいたことが予想されるので，生存者バイアスが生じていると考えられる。*POS データも同様。*

例２．ドイツを愛するハズレないタコ

2010 年に一匹のタコが有名になったそうです。FIFA ワールドカップ 2010 南アフリカ大会で，ドイツ代表の試合８つの勝敗をすべて的中させたというのです。餌と一緒に置いた２つの国旗のどちらへ行くかで決めたそうです。

［**考えてみよう２**］８つの試合の勝敗をすべて的中させる確率はどれくらいなのか，計算して分数で求めてみましょう。（引き分けは考えません。）

勝つか負けるかのどちらかとすると，的中の確率は$\frac{1}{2}$ですから，８つ連続で$\left(\frac{1}{2}\right)^8=\frac{1}{2^8}=\frac{1}{256}$
つまり，256 回に 1 回的中する確率になります。

●これがとても珍しいことと言ってよいのか…というのは前提条件によるところがある。ワールドカップの勝敗は世界中の多くの人が関心を寄せ，同じように動物に勝敗を占わせていたのが（　２５６　）匹以上いたとすると，少なくとも 1 匹以上は確率的に的中するはずである。このとき，多くの（　　**黙って隠れている**　　）**存在がある**，ならば確率的に珍しいとは言えなくなる。これも一つの**バイアス**と言える。

例３．食い違った２つの大統領選挙予想

1936 年のアメリカ大統領選挙においては，民主党ルーズベルト大統領と共和党ランドン候補が争っていました。このとき，２つの世論調査の結果も対立しました。

①「リテラリー・ダイジェスト誌」…過去 5 回の大統領選挙結果を当てており，サンプル数は 200 万人。ランドン候補が 57%の支持で有利と発表。

②「ギャラップ（当時はアメリカ世論研究所）」…当時は新興の調査会社であり，サンプル数は 3000 人。ルーズベルト大統領が 54%の支持で有利と発表。

結果はルーズベルト大統領の圧勝となり，世間を驚かせました。

［**考えてみよう３**］当時は①の調査結果が支持されていました。そのように信頼性が高いと思われていた理由を 2 点，簡潔に「～ので，～と思われていたから」という文で説明してみましょう。

- *過去 5 回の結果を当てた実績があるので，今回も当たる*　　と思われていたから
- *サンプル数が一方の約 700 倍なので，より正確な結果だ*　　と思われていたから

●サンプル選びの違い…①と②ではサンプルの選び方に違いがあった。
①自社の雑誌読者，電話や自動車を所有する人を対象に得たサンプル
②都市の男女，農村の男女，中間層の男女…等の**層に細分化**し有権者数に合わせたサンプル
①のサンプルは結果として富裕層を中心に選んだことになり，それまでは母集団とあまり違わなかったものが，経済状況等の社会情勢の変化によって当てはまらなくなったと考えられる。②のようなサンプルの集め方は（　　**層別サンプリング**　　）という。

●サンプル選びの方法…人が選ぶと頼みやすさ等で偏りがでることから，コンピュータを用いた（　　**ランダムサンプリング**　　）**法**もしくは（　　**無作為抽出**　　）**法**と呼ばれる方法が利用されている。その一つが無作為に電話番号を生成して電話して調査を依頼する**RDD（Random Digit Dialing）法**であるが，これも間違いない方法とは言い切れない。