# 1時限で理解する統計の基礎

応用情報処理II
2017/12/1
講師：新居雅行

# 今日の目的

- 統計は難しい、けど知らずにパソコンに向かってやり方だけ勉強しても仕方ない
- だけど、まじめに勉強する機会も少ない
- まじめに勉強することを勧めるが、最低限の知識を今日の1時限で詰め込む

# 統計とは

- 過去に起こった事実を
  - あくまで事実を求めるというスタンスが基本にある
- 数値的に評価するもの
  - 定性的に評価するものではない
- すなわち、現象や実態を、客観的に判断するためのよりどころとしての統計がある

# 統計は数学の1分野である

- 数字を求め、数字を評価の基礎とする
- 数字を求めるためには計算が必要
  - しかしながら、鶴亀算じゃあるまいし、手順化は手詰まりになる
- そこで、数式をベースにした一般化が図られる
  - 微積分（解析学）の基礎の上にあるので、それを知らないと厳しい面もある
  - 数学の強みと大変な面は、いずれも「一般化」されているという点である

# 数学と統計学の違い

- イコールは、実は＝ではない
  - 例えば、平均値=合計÷個数
  - このイコールは何を意味するか？
  - 公式にあてはめて求めた数値は、実は推定値であるというのが一般的なスタンス
- 真の意味でのイコールではない
- 数学的な意味ではイコールでかまわない
  - 計算結果を求めるという意味ではイコールである

# 確率と統計

- 確率は、どちらかというと未来に起こるできごとを、数学的に推定するといった世界
  - したがって、起こってもいないことをあれこれ言うというこれも不思議な世界
- ただし、確率を求めるよりどころは統計にあるというのが一般的
  - 確率論によるモデル化をベースに統計がある
  - 数学的な意味付けは、確率の考え方が基本にある

# 確率の例

- サイコロを2つ振って、同じ目が出る確率
  - サイコロの6面は、同じ確率（1/6）で出てくる
  - 組み合わせは、6×6=36通り
  - 同じ目が出るのは全部で6通り
  - 従って、1/6=0.16…(17%)
- 確率の数値は解釈が必要
  - たとえば、100回振り、同じ目が出る回数をカウントする
  - いつも、17回とは限らない、13回かもしれないが、20回かもしれない
  - 100回の試行をたくさん行うと、恐らく17回の場合がいちばん多くなるはず

# 統計の非常に重要な概念

- 母集団とサンプル
  - 測定値は元データなのか、元データの一部を取り出した者なのか？
- 事象は確率的に発生する
  - 一見ランダムに見えても、一定の統計モデルに従う
  - 言い換えれば、統計モデルに合致する部分を見つける
- 平均
  - これを理解できれば統計は制覇したものと同じ！というのは言い過ぎかも
  - しかし、あまりに意味が深く、勉強して、勉強して、行き着いたのは平均だった

# 平均

- 求め方はもう説明は必要ないでしょう
  - 合計を個数で割る
- 平均の意味は
  - 誤差がいちばん少ない数
- 非常に誤解しやすい点
  - 単に計算方法を知っているのは何の意味もない。たとえば、1人の人の身長と体重の平均値は何か意味はあるか？
  - 統計のポイントになるが、常に「意味」「背景」を頭にいれておくことが大切

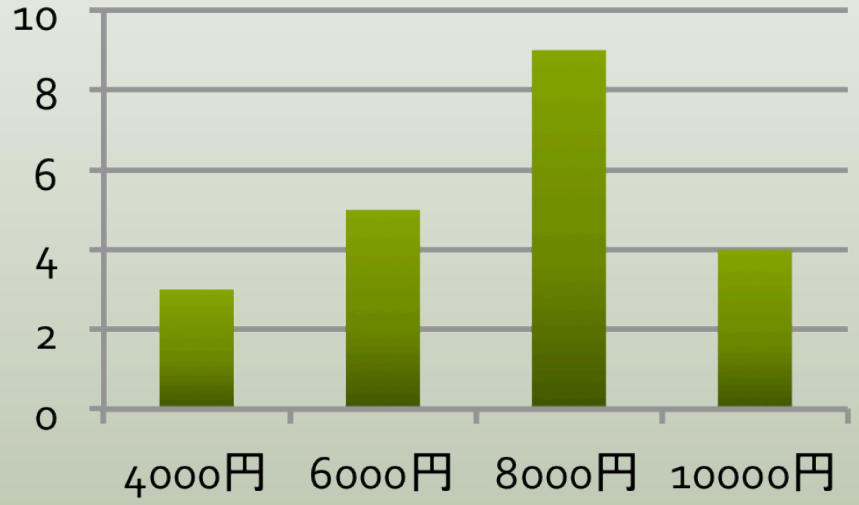
# 平均の求め方

- 身長が167,158,173,159の平均値
  - もちろん、(167+158+173+159)÷4=164.25
- ちょっと考えよう
  - 164.25の意味
  - この4人の中には、164.25という身長のやつはいないぞ
  - 実は「比較」において意味がある数値。比較の方法は検定などとも呼ばれている

# 平均を求める方法

- こんなデータがあるとする
  - 以下の計算式で平均値を求められる
  - ((4000×3)+(6000×5)+(8000×9)+(10000×4))÷(3+5+9+4)=7333.33…
  - 要はヒストグラム

| | |
|---|---|
| 4000円 | 3人 |
| 6000円 | 5人 |
| 8000円 | 9人 |
| 10000円 | 4人 |



# 分散

- データの散らばり具合
  - 平均値との差を2乗した値は、はずれ値になるほど大きな数値になる
  - しかも2乗するので、はずれればはずれるほど、その傾向が増幅される
- その平均値をとって「分散」と呼ぶ
- 前のプレゼンの図の場合
  - ((4000-7333)^2*3+(6000-7333)^2*5+(8000-7333)^2*9+(10000-7333)^2*4)/21 = 3555555.5…

# 標準偏差

- 分散の単位は、元データの2乗になっているので、単位も2乗になる
  - だから、そのルートを取れば単位は揃う
- 結果的に散らばり具合を示す指標としての標準偏差が求められる
- 前のプレゼンの図の場合
  - 3555555のルート=1885.618…

# サンプリングと母集団

- 同じ統計値でも場面で異なる
  - 母集団：クラスの試験の成績
  - サンプリング：クラスの試験の成績はその学校の学力を示すものだ
- サンプリング結果から、母集団の統計値を推定する
  - 平均値の推定値＝サンプルの平均値
  - 分散の推定値＝ちょっと式が変わる→これを「標本標準偏差」と呼ぶ

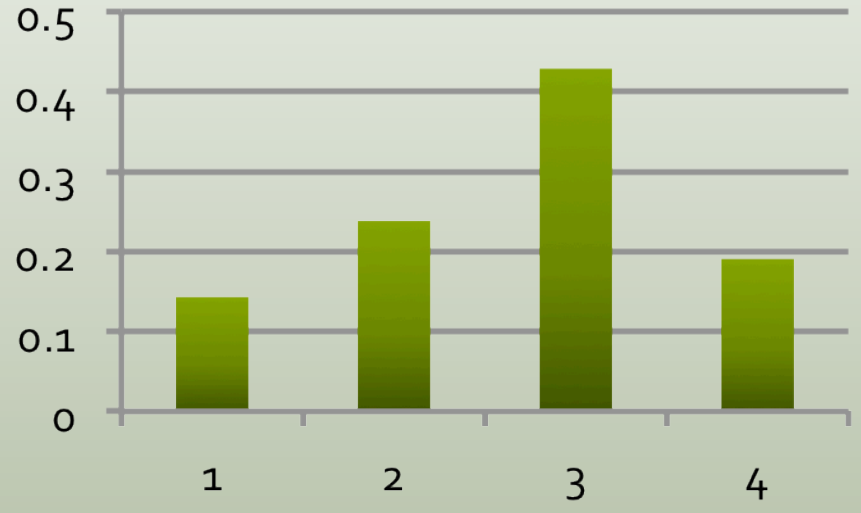
# 標本標準偏差

- 平均値との差の2乗値を、（個数-1）で割る
  - そのルートが標本標準偏差
  - つまり、少し大きくなる
- 数学的には証明などができるのだが、考え方として、ばらつきは広がる可能性があると考える
- 前のプレゼンの図の場合
  - ((4000-7333)^2*3+(6000-7333)^2*5+(8000-7333)^2*9+(10000-7333)^2*4)/(21-1) の平方根 ≈ 1932

# 確率分布

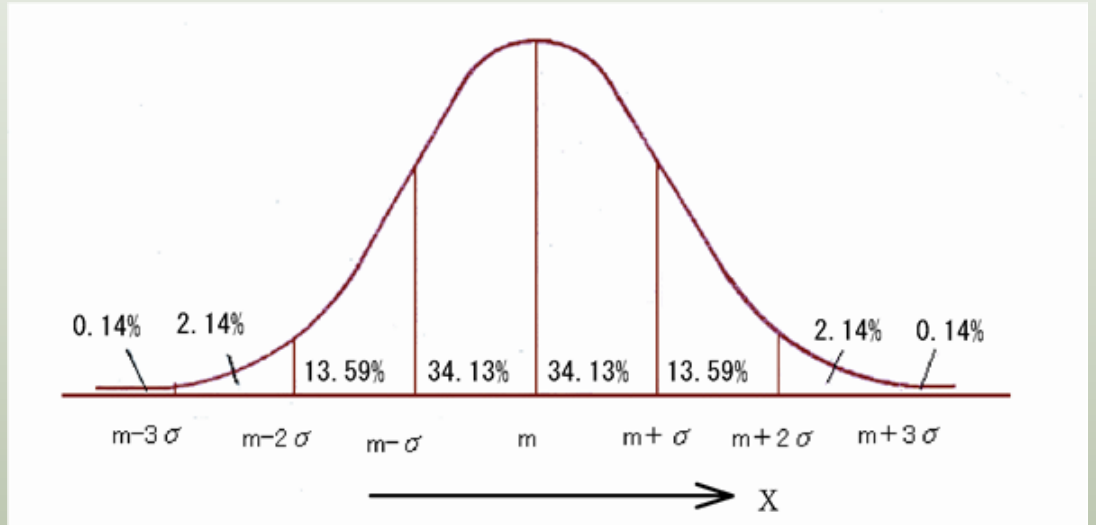
- 縦軸に確率を取る
- 数学的には関数で表現される
- 要はヒストグラム

| | | |
|---|---|---|
| 4000円 | 3人 | 3/21=14% |
| 6000円 | 5人 | 5/21=24% |
| 8000円 | 9人 | 9/21=43% |
| 10000円 | 4人 | 4/21=19% |



# 正規分布

- 確率分布の代表的な形式
  - 平均値を中心に分布は左右対称になり、平均値から離れるほど頻度が低下する
- 偶然が重なることによって、正規分布になるとされている




正規分布

http://www.weblio.jp/content/正規分布

# 正規分布であるなら

- もし、測定値が正規分布であると言えるなら
  - 計算された平均値と標本標準偏差は、母集団の平均値と標準偏差の最も確かな推定値である
  - 平均値±標本標準偏差の間に、測定値の34.1×2=68.2%のものが含まれるだろう
  - 平均値±2×標本標準偏差の間に、測定値の約96%のものが含まれるだろう
  - 前の例：3469～11198の範囲に96%のデータが含まれるだろう
- 問題は、本当に正規分布するのかどうか？

# 推定と検定

- 推定
  - 統計量をもとに、ある確率で当たるという前提をおいて、区間などを求める
- 検定
  - 実験や調査の「結果」に使われることがよくある
  - 統計値（平均値と標準偏差）のペアに対し、仮説として「2つの測定値は等しい」を立てる
  - その仮説は間違いであるという場合において意味がある（帰無仮説）
  - 2つの測定値は「同じではない」ということを「違っている」とみなすのがある意味では検定の核心である

# 各種の統計解析

- 分散分析
- 回帰分析
- 多変量解析
- これらは、データの傾向を語るのに使われる

# 統計的な手法

- 尺度の問題
  - 測定値が加算可能かを考える
  - 加算できない数値とは？
- 確率の問題か、必ず起こるという問題か？
- ここを間違えると、あらゆる統計的な手法は意味がなくなる

# 統計の勉強方法

- とにかくなにか1冊は破読すること
- 必ず、サンプルのデータを自分の手で計算をしてみること
- どんな複雑な解析手法でも、一度は手作業で解くこと。それから、コンピュータを使うように
- 法則や手順がほんとうに適用可能なのかを考えられるようになることが必要